Text as Data

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

September 25th, 2014

A General Framework for the Class

A focus on tasks

A General Framework for the Class

A focus on tasks

1) Decide on some objective to accomplish

A General Framework for the Class

A focus on tasks

1) Decide on some objective to accomplish
    - Classify

A General Framework for the Class

A focus on tasks

1) Decide on some objective to accomplish
    - Classify
    - Cluster

A General Framework for the Class

A focus on tasks

1) Decide on some objective to accomplish
   - Classify
   - Cluster
   - Predict

A General Framework for the Class

A focus on tasks

1) Decide on some objective to accomplish
   - Classify
   - Cluster
   - Predict
   - Describe

A General Framework for the Class

A focus on tasks

1) Decide on some objective to accomplish
   - Classify
   - Cluster
   - Predict
   - Describe
   - Measure covariance, discover latent structure, find nearest neighbor, ...

A General Framework for the Class

2) Use an objective function $\rightsquigarrow$ measure performance at task

A General Framework for the Class

2) Use an objective function $\rightsquigarrow$ measure performance at task
   Suppose we have data $\boldsymbol{X} \in \mathcal{X}$ and parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

A General Framework for the Class

2) Use an objective function $\rightsquigarrow$ measure performance at task
   Suppose we have data $\boldsymbol{X} \in \mathcal{X}$ and parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$f : \mathcal{X} \times \boldsymbol{\Theta} \rightarrow \Re$$

A General Framework for the Class

2) Use an objective function $\rightsquigarrow$ measure performance at task
   Suppose we have data $\boldsymbol{X} \in \mathcal{X}$ and parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$f : \mathcal{X} \times \boldsymbol{\Theta} \rightarrow \Re$$
$$f(\boldsymbol{X}, \theta)$$

A General Framework for the Class

2) Use an objective function $\rightsquigarrow$ measure performance at task
   Suppose we have data $\boldsymbol{X} \in \mathcal{X}$ and parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$f : \mathcal{X} \times \boldsymbol{\Theta} \to \Re$$
$$f(\boldsymbol{X}, \boldsymbol{\theta})$$

- Derived from task/used for statistical properties$\rightsquigarrow$ minimize sum of squared residuals

A General Framework for the Class

2) Use an objective function $\rightsquigarrow$ measure performance at task
   Suppose we have data $\boldsymbol{X} \in \mathcal{X}$ and parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$
\begin{aligned}
f &: \mathcal{X} \times \boldsymbol{\Theta} \rightarrow \Re \\
f&(\boldsymbol{X}, \boldsymbol{\theta})
\end{aligned}
$$

- Derived from task/used for statistical properties $\rightsquigarrow$ minimize sum of squared residuals
- Formalization of intuition about "good" performance $\rightsquigarrow$ k-means clustering

A General Framework for the Class

2) Use an objective function $\rightsquigarrow$ measure performance at task
   Suppose we have data $\boldsymbol{X} \in \mathcal{X}$ and parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$f : \mathcal{X} \times \boldsymbol{\Theta} \rightarrow \Re$$
$$f(\boldsymbol{X}, \boldsymbol{\theta})$$

- Derived from task/used for statistical properties$\rightsquigarrow$ minimize sum of squared residuals
- Formalization of intuition about "good" performance $\rightsquigarrow$ k-means clustering
- Data generating process

A general Framework for the Class

3) Method for optimizing objective function.

A general Framework for the Class

3) Method for optimizing objective function.
   Find $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ such that

A general Framework for the Class

3) Method for optimizing objective function.
   Find $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ such that

$$f(\boldsymbol{X}, \boldsymbol{\theta}^*)$$

A general Framework for the Class

3) Method for optimizing objective function.
   Find $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ such that

$$f(\boldsymbol{X}, \boldsymbol{\theta}^*)$$

   is a maximum (minimum)

A general Framework for the Class

3) Method for optimizing objective function.
   Find $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ such that

$$f(\boldsymbol{X}, \boldsymbol{\theta}^*)$$

   is a maximum (minimum)
   - Analytic methods (Calculus)

A general Framework for the Class

3) Method for optimizing objective function.
   Find $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ such that

$$f(\boldsymbol{X}, \boldsymbol{\theta}^*)$$

   is a maximum (minimum)

   - Analytic methods (Calculus)
   - Computational methods

Today: Using (Bayesian) Statistics to Obtain Objective Functions

- Encode assumptions in data generating process $\rightsquigarrow$ hierarchical model
- Assume parameters and data are random variables
- Conditional probability statement $\rightsquigarrow$ objective function
- Use computational tools to optimize objective function

Today: Using (Bayesian) Statistics to Obtain Objective Functions

Plan of Attack:

1) Write out any joint density function as conditional relationship
2) Show how this can be an objective function even if you've never taken likelihood/Bayesian/...
3) Discuss how to computationally optimize

Joint Distributions of Random Variables

Definition

Suppose that we have random variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$*. We will say that* $\boldsymbol{X}$ *is a jointly continuous random variable if for all* $\boldsymbol{X} \in \Re^K$ *there exists a function* $f : \Re^K \rightarrow \Re$ *such that for all* $C \subset \Re^K$*,*

$$P(\boldsymbol{X} \in C) = \iint \ldots \iint_{(\boldsymbol{X}) \in C} f(\boldsymbol{x}) d\boldsymbol{X}$$

- A joint density $f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_K)$ encodes information about the behavior of the random variable $\boldsymbol{X}$

Marginal Distribution

Definition

Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$ is a jointly continuous random variable. Define $f_{X_j}(x)$ as the marginal *probability density function for X_j,*

$$
\begin{aligned}
f_{X_j}(x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\boldsymbol{x})dx_1 dx_2 \ldots dx_{j-1}dx_{j+1}\ldots dx_{K-1}dx_K \\
&= \int_{\Re^{K-1}} f(\boldsymbol{x})d\boldsymbol{X}_{-j}
\end{aligned}
$$

- To obtain the marginal distribution, $f_{X_j}(x)$ we integrate over all dimensions but j

Conditional Distributions and Independence of Random Variables

Definition

Suppose $\boldsymbol{X}$ *is a jointly continuous random variable. Define* $f_{\boldsymbol{X}_{-j}|X_j}(\boldsymbol{x})$ *as the conditional density function,*

$$f_{\boldsymbol{X}_{-j}|X_j}(\boldsymbol{x}_{-j}|x_j) = \frac{f(x_1, x_2, \ldots, x_K)}{f_{X_j}(x_j)}$$

Conditional Distributions and Independence of Random Variables

Definition

Suppose $\boldsymbol{X}$ *is a jointly continuous random variable. Define* $f_{\boldsymbol{X}_{-j}|X_j}(\boldsymbol{x})$ *as the conditional density function,*

$$f_{\boldsymbol{X}_{-j}|X_j}(\boldsymbol{x}_{-j}|x_j) = \frac{f(x_1, x_2, \ldots, x_K)}{f_{X_j}(x_j)}$$

Two random variables X_1 *and* X_2 *are independent if*

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

Conditional Independence of Random Variables

Definition

Two random variables X_1 and X_2 are conditionally independent given X_3 if

$$f_{(X_1,X_2)|X_3}(x_1, x_2|x_3) = f_{X_1|X_3}(x_1|x_3)f_{X_2|X_3}(x_2|x_3)$$

Conditional Independence of Random Variables

Definition

Two random variables X_1 and X_2 are conditionally independent given X_3 if

$$f_{(X_1,X_2)|X_3}(x_1, x_2|x_3) \quad = \quad f_{X_1|X_3}(x_1|x_3)f_{X_2|X_3}(x_2|x_3)$$

Equivalently,

$$f_{X_1|X_2,X_3}(x_1|x_2, x_3) \quad = \quad f_{X_1|X_3}(x_1|x_3)$$

Joint Density as Conditional Relationship

Theorem

Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$ *is a jointly continuous random variable. Then*

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(x_1, x_2, \ldots, x_K) \\
&= f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_2 X_1}(x_3|x_2, x_1) \ldots f_{X_K|X_1 \ldots X_{K-1}}(x_K|x_1, \ldots, x_{K-1})
\end{aligned}
$$

Joint Density as Conditional Relationship

Theorem

Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$ *is a jointly continuous random variable. Then*

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(x_1, x_2, \ldots, x_K) \\
&= f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_2 X_1}(x_3|x_2, x_1) \ldots f_{X_K|X_1 \ldots X_{K-1}}(x_K|x_1, \ldots, x_{K-1})
\end{aligned}
$$

- We can write joint distributions as a product of conditional distributions

Joint Density as Conditional Relationship

Theorem

Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$ *is a jointly continuous random variable. Then*

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(x_1, x_2, \ldots, x_K) \\
&= f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_2 X_1}(x_3|x_2, x_1) \ldots f_{X_K|X_1 \ldots X_{K-1}}(x_K|x_1, \ldots, x_{K-1})
\end{aligned}
$$

- We can write joint distributions as a product of conditional distributions
- If there are conditional independences in density we can simplify $\rightsquigarrow$ simplify some conditional expressions

Beta-Binomial Model

Definition

Suppose Y is a continuous random variable with $Y \in [0,1]$ and pdf of Y given by

$$f(y) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1}(1-y)^{\alpha_2 - 1}$$

Then we will say Y is a Beta distribution with parameters α_1 and α_2. Equivalently,

$$Y \sim Beta(\alpha_1, \alpha_2)$$

Beta-Binomial Model

Definition

Suppose Y is a continuous random variable with $Y \in [0,1]$ and pdf of Y given by

$$f(y) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1}(1-y)^{\alpha_2 - 1}$$

Then we will say Y is a Beta distribution with parameters α_1 and α_2. Equivalently,

$$Y \sim Beta(\alpha_1, \alpha_2)$$

- Beta is a distribution on proportions

Beta-Binomial Model

Definition

Suppose Y *is a continuous random variable with* $Y \in [0, 1]$ *and pdf of* Y *given by*

$$f(y) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1}(1 - y)^{\alpha_2 - 1}$$

Then we will say Y *is a Beta distribution with parameters* α_1 *and* α_2. *Equivalently,*

$$Y \sim Beta(\alpha_1, \alpha_2)$$

- Beta is a distribution on proportions
- Beta is a special case of the Dirichlet distribution

Beta-Binomial Model

Definition

Suppose Y is a continuous random variable with $Y \in [0,1]$ and pdf of Y given by

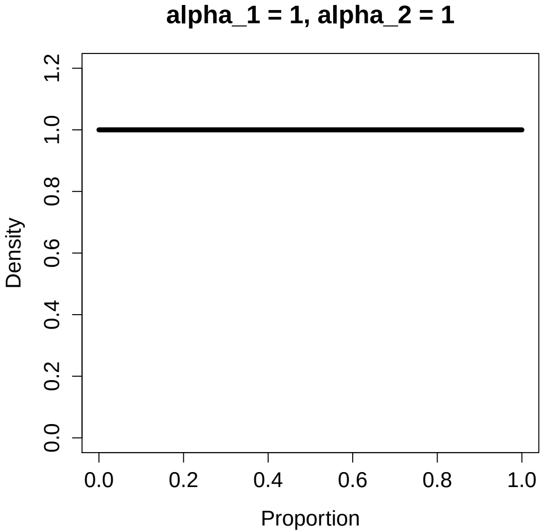
$$f(y) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1}(1-y)^{\alpha_2 - 1}$$

Then we will say Y is a Beta distribution with parameters α_1 and α_2. Equivalently,

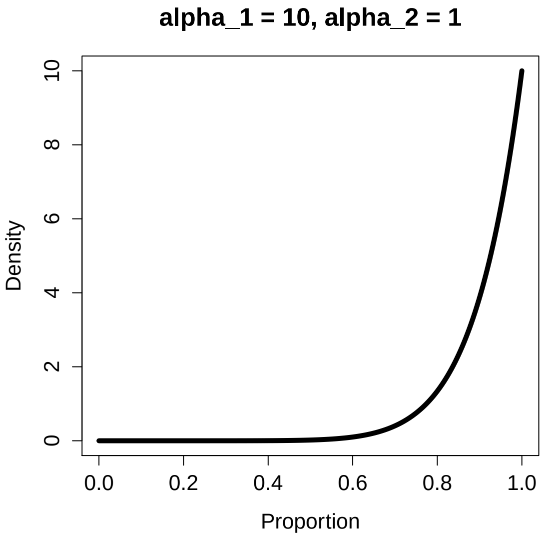
$$Y \sim Beta(\alpha_1, \alpha_2)$$

- Beta is a distribution on proportions
- Beta is a special case of the Dirichlet distribution
- $E[Y] = \frac{\alpha_1}{\alpha_1 + \alpha_2}$

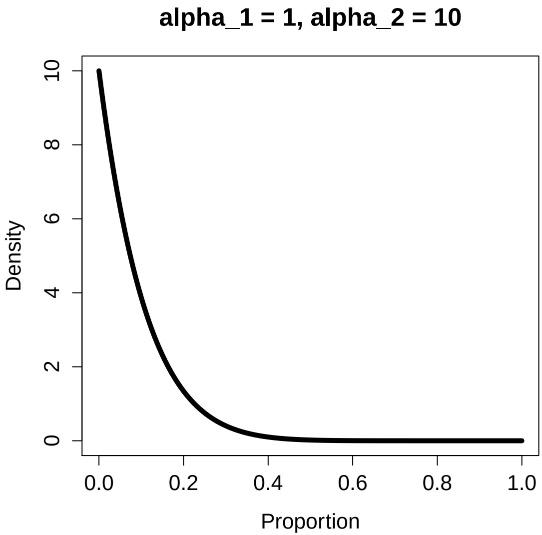
Beta Distribution



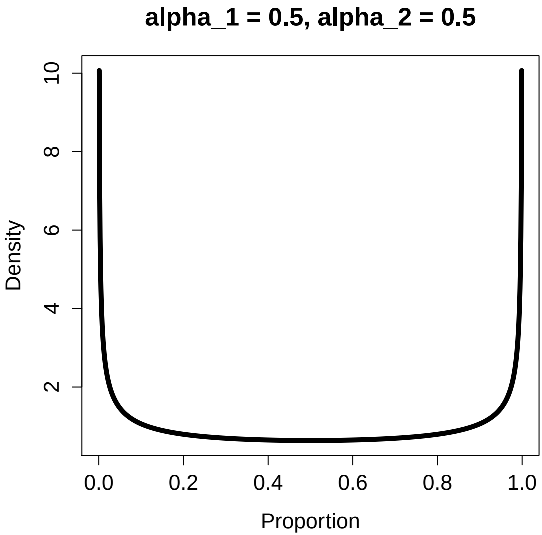
Beta Distribution



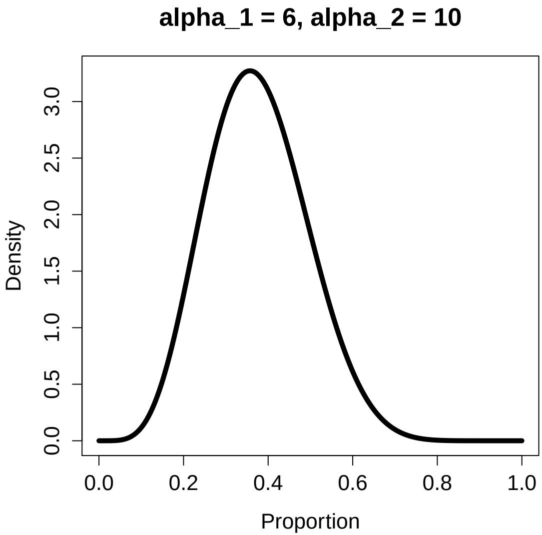
Beta Distribution



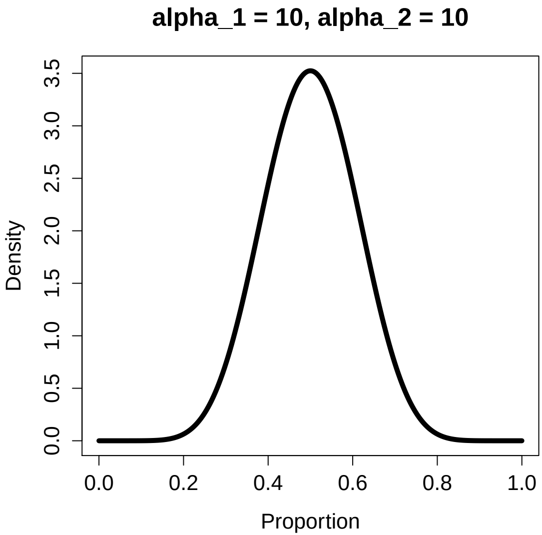
Beta Distribution



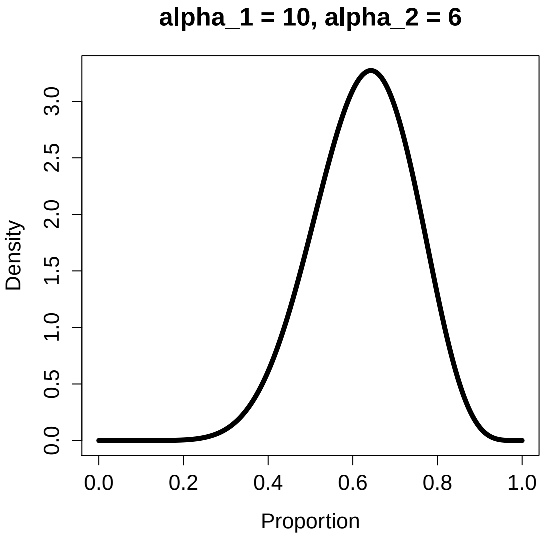
Beta Distribution



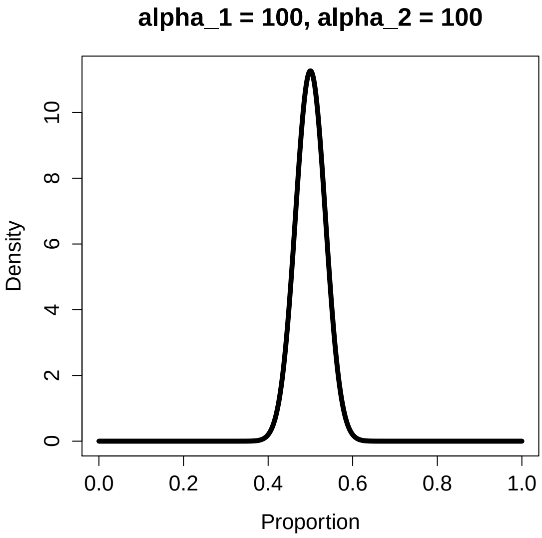
Beta Distribution



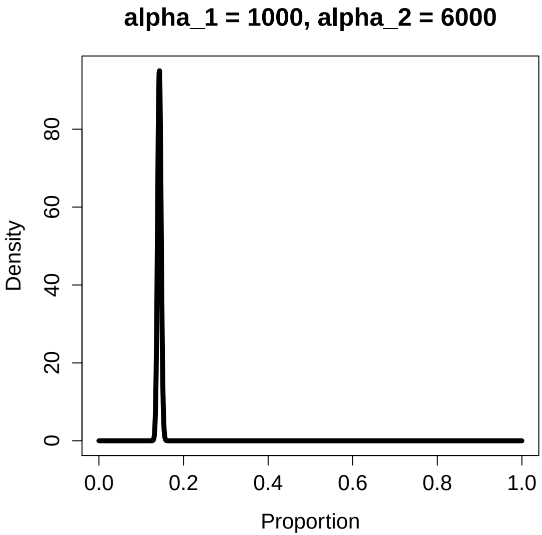
Beta Distribution



Beta Distribution



Beta Distribution



Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements

Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements

We want to infer π, we observe incidence in a statement

Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements

We want to infer π, we observe incidence in a statement

$$\pi \sim \text{Beta}(\alpha_1, \alpha_2)$$

Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements
We want to infer π, we observe incidence in a statement

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y|\pi &\sim \text{Bernoulli}(\pi) \end{aligned}$$

Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements

We want to infer π, we observe incidence in a statement

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y|\pi &\sim \text{Bernoulli}(\pi) \end{aligned}$$

We can write $p(\pi, y|\alpha_1, \alpha_2)$ as

Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements

We want to infer π, we observe incidence in a statement

$$\begin{aligned} \pi & \sim \text{Beta}(\alpha_1, \alpha_2) \\ Y|\pi & \sim \text{Bernoulli}(\pi) \end{aligned}$$

We can write $p(\pi, y|\alpha_1, \alpha_2)$ as

$$p(\pi, y|\alpha_1, \alpha_2)$$

Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements

We want to infer π, we observe incidence in a statement

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y|\pi &\sim \text{Bernoulli}(\pi) \end{aligned}$$

We can write $p(\pi, y|\alpha_1, \alpha_2)$ as

$$p(\pi, y|\alpha_1, \alpha_2) = p(\pi|\alpha_1, \alpha_2)p(y|\pi, \alpha_1, \alpha_2) = p(\pi|\alpha_1, \alpha_2)p(y|\pi)$$

Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements

We want to infer π, we observe incidence in a statement

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y|\pi &\sim \text{Bernoulli}(\pi) \end{aligned}$$

We can write $p(\pi, y|\alpha_1, \alpha_2)$ as

$$\begin{aligned} p(\pi, y|\alpha_1, \alpha_2) &= p(\pi|\alpha_1, \alpha_2)p(y|\pi, \alpha_1, \alpha_2) = p(\pi|\alpha_1, \alpha_2)p(y|\pi) \\ &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}\pi^{y}(1-\pi)^{1-y} \end{aligned}$$

Beta-Binomial Model

Suppose we are interested in the use of the word "Obamacare" in a legislator's statements
We want to infer π, we observe incidence in a statement

$$\begin{aligned}
\pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\
Y|\pi &\sim \text{Bernoulli}(\pi)
\end{aligned}$$

We can write $p(\pi, y|\alpha_1, \alpha_2)$ as

$$\begin{aligned}
p(\pi, y|\alpha_1, \alpha_2) &= p(\pi|\alpha_1, \alpha_2)p(y|\pi, \alpha_1, \alpha_2) = p(\pi|\alpha_1, \alpha_2)p(y|\pi) \\
&= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}\pi^{y}(1-\pi)^{1-y} \\
&= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\alpha_1 + y - 1}(1-\pi)^{\alpha_2 + 1 - y - 1}
\end{aligned}$$

Beta-Binomial Model

Suppose now we observe n statements

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y_i|\pi &\sim \text{Bernoulli}(\pi) \text{ for } i = 1, \ldots, n \end{aligned}$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$

Beta-Binomial Model

Suppose now we observe n statements

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y_i|\pi &\sim \text{Bernoulli}(\pi) \text{ for } i = 1, \ldots, n \end{aligned}$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$
We can write $p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2)$ as

Beta-Binomial Model

Suppose now we observe n statements

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y_i|\pi &\sim \text{Bernoulli}(\pi) \text{ for } i = 1, \ldots, n \end{aligned}$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$
We can write $p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2)$ as

$$p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2)$$

Beta-Binomial Model

Suppose now we observe n statements

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y_i|\pi &\sim \text{Bernoulli}(\pi) \text{ for } i = 1, \ldots, n \end{aligned}$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$
We can write $p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2)$ as

$$p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2) = p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi)$$

Beta-Binomial Model

Suppose now we observe n statements

$$\begin{aligned} \pi & \sim \text{Beta}(\alpha_1, \alpha_2) \\ Y_i|\pi & \sim \text{Bernoulli}(\pi) \text{ for } i = 1, \ldots, n \end{aligned}$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$
We can write $p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2)$ as

$$\begin{aligned} p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2) & = p(\pi|\alpha_1, \alpha_2) p(\boldsymbol{y}|\pi) \\ & = p(\pi|\alpha_1, \alpha_2) \prod_{i=1}^{n} p(y_i|\pi) \end{aligned}$$

Beta-Binomial Model

Suppose now we observe n statements

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\ Y_i|\pi &\sim \text{Bernoulli}(\pi) \text{ for } i = 1, \ldots, n \end{aligned}$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$
We can write $p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2)$ as

$$\begin{aligned} p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2) &= p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi) \\ &= p(\pi|\alpha_1, \alpha_2)\prod_{i=1}^{n} p(y_i|\pi) \\ &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1}\prod_{i=1}^{n}\pi^{y_i}(1-\pi)^{1-y_i} \end{aligned}$$

Beta-Binomial Model

Suppose now we observe n statements

$$
\begin{aligned}
\pi &\sim \text{Beta}(\alpha_1, \alpha_2) \\
Y_i|\pi &\sim \text{Bernoulli}(\pi) \text{ for } i = 1, \ldots, n
\end{aligned}
$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$
We can write $p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2)$ as

$$
\begin{aligned}
p(\pi, \boldsymbol{y}|\alpha_1, \alpha_2) &= p(\pi|\alpha_1, \alpha_2) p(\boldsymbol{y}|\pi) \\
&= p(\pi|\alpha_1, \alpha_2) \prod_{i=1}^{n} p(y_i|\pi) \\
&= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\alpha_1 - 1}(1-\pi)^{\alpha_2 - 1} \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \\
&= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\sum_{i=1}^{n} y_i + \alpha_1 - 1}(1-\pi)^{n + \alpha_2 - \sum_{i=1}^{n} y_i - 1}
\end{aligned}
$$

Bayes' Theorem for Continuous Random Variables

Theorem

Suppose we have jointly continuous random variables X_1 and X_2. Then,

$$f_{X_1|X_2}(x_1|x_2) \quad = \quad \frac{f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1)}{f_{X_2}(x_2)}$$

Beta-Binomial Model

We observe $\boldsymbol{y}$ and we want to learn about π

Beta-Binomial Model

We observe $\boldsymbol{y}$ and we want to learn about π
Condition on data, describe function of π.

Beta-Binomial Model

We observe $\boldsymbol{y}$ and we want to learn about π
Condition on data, describe function of π.

$$p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2) = \frac{p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi, \alpha_1, \alpha_2)}{p(\boldsymbol{y})}$$

Beta-Binomial Model

We observe $\boldsymbol{y}$ and we want to learn about π
Condition on data, describe function of π.

$$
\begin{aligned}
p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2) &= \frac{p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi, \alpha_1, \alpha_2)}{p(\boldsymbol{y})} \\
&\propto p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi)
\end{aligned}
$$

Beta-Binomial Model

We observe $\boldsymbol{y}$ and we want to learn about π
Condition on data, describe function of π.

$$
\begin{aligned}
p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2) &= \frac{p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi, \alpha_1, \alpha_2)}{p(\boldsymbol{y})} \\
&\propto p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi) \\
&\propto \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\sum_{i=1}^{n} y_i + \alpha_1 - 1}(1-\pi)^{n+\alpha_2 - \sum_{i=1}^{n} y_i - 1}
\end{aligned}
$$

Beta-Binomial Model

We observe $\boldsymbol{y}$ and we want to learn about π

Condition on data, describe function of π.

$$
\begin{aligned}
p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2) &= \frac{p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi, \alpha_1, \alpha_2)}{p(\boldsymbol{y})} \\
&\propto p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi) \\
&\propto \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\pi^{\sum_{i=1}^{n} y_i + \alpha_1 - 1}(1 - \pi)^{n + \alpha_2 - \sum_{i=1}^{n} y_i - 1}
\end{aligned}
$$

Defines a function of π, which we can use to describe the data.

Beta-Binomial Model

We observe $\boldsymbol{y}$ and we want to learn about π

Condition on data, describe function of π.

$$\begin{aligned} p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2) &= \frac{p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi, \alpha_1, \alpha_2)}{p(\boldsymbol{y})} \\ &\propto p(\pi|\alpha_1, \alpha_2)p(\boldsymbol{y}|\pi) \\ &\propto \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\sum_{i=1}^n y_i + \alpha_1 - 1}(1-\pi)^{n + \alpha_2 - \sum_{i=1}^n y_i - 1} \end{aligned}$$

Defines a function of π, which we can use to describe the data.

Optimize $\rightsquigarrow$ analytically or computationally.

Theorem

Suppose $f : \Re^K \rightarrow \Re_+$. *If* $\boldsymbol{x}^*$ *maximizes* $\log(f(x))$, *then* $\boldsymbol{x}^*$ *maximizes* $f(x)$.

Theorem

Suppose $f : \Re^K \rightarrow \Re_+$. *If* $\boldsymbol{x}^*$ *maximizes* $\log(f(x))$, *then* $\boldsymbol{x}^*$ *maximizes* $f(x)$.

$$\log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right)$$

Theorem

Suppose $f : \Re^K \rightarrow \Re_{+}$. If $\boldsymbol{x}^*$ maximizes $\log(f(x))$, then $\boldsymbol{x}^*$ maximizes $f(x)$.

$$
\begin{aligned}
\log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right) &= \left(\sum_{i=1}^{n} y_i + \alpha_1 - 1\right)\log \pi \\
&\quad + (n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)(1 - \log \pi) + c
\end{aligned}
$$

Theorem

Suppose $f : \Re^K \rightarrow \Re_+$. *If* $\boldsymbol{x}^*$ *maximizes* $\log(f(x))$, *then* $\boldsymbol{x}^*$ *maximizes* $f(x)$.

$$\begin{aligned}
\log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right) &= \left(\sum_{i=1}^{n} y_i + \alpha_1 - 1\right)\log \pi \\
&\quad + (n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)(1 - \log \pi) + c \\
\frac{\partial \log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right)}{\partial \pi} &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{\pi} - \frac{(n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)}{1 - \pi}
\end{aligned}$$

Theorem

Suppose $f : \Re^K \rightarrow \Re_{+}$. *If* $\boldsymbol{x}^*$ *maximizes* $\log(f(x))$, *then* $\boldsymbol{x}^*$ *maximizes* $f(x)$.

$$
\begin{aligned}
\log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right) &= \left(\sum_{i=1}^{n} y_i + \alpha_1 - 1\right)\log \pi \\
&\quad + (n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)(1 - \log \pi) + c \\
\frac{\partial \log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right)}{\partial \pi} &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{\pi} - \frac{(n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)}{1 - \pi} \\
0 &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{\pi^*} - \frac{n + \alpha_2 - \sum_{i=1}^{n} y_i - 1}{1 - \pi^*}
\end{aligned}
$$

Theorem

Suppose $f : \Re^K \rightarrow \Re_{+}$. *If* $\boldsymbol{x}^*$ *maximizes* $\log(f(x))$, *then* $\boldsymbol{x}^*$ *maximizes* $f(x)$.

$$
\begin{aligned}
\log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right) &= \left(\sum_{i=1}^{n} y_i + \alpha_1 - 1\right)\log \pi \\
&\quad + (n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)(1 - \log \pi) + c \\
\frac{\partial \log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right)}{\partial \pi} &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{\pi} - \frac{(n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)}{1 - \pi} \\
0 &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{\pi^*} - \frac{n + \alpha_2 - \sum_{i=1}^{n} y_i - 1}{1 - \pi^*} \\
\pi^* &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{n + \alpha_1 + \alpha_2 - 2}
\end{aligned}
$$

Theorem

Suppose $f : \Re^K \rightarrow \Re_+$. *If* $\boldsymbol{x}^*$ *maximizes* $\log(f(x))$, *then* $\boldsymbol{x}^*$ *maximizes* $f(x)$.

$$
\begin{aligned}
\log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right) &= \left(\sum_{i=1}^{n} y_i + \alpha_1 - 1\right) \log \pi \\
& \quad + (n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)(1 - \log \pi) + c \\
\frac{\partial \log\left(p(\pi|\boldsymbol{y}, \alpha_1, \alpha_2)\right)}{\partial \pi} &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{\pi} - \frac{(n + \alpha_2 - \sum_{i=1}^{n} y_i - 1)}{1 - \pi} \\
0 &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{\pi^*} - \frac{n + \alpha_2 - \sum_{i=1}^{n} y_i - 1}{1 - \pi^*} \\
\pi^* &= \frac{\sum_{i=1}^{n} y_i + \alpha_1 - 1}{n + \alpha_1 + \alpha_2 - 2}
\end{aligned}
$$

Second derivative test $\rightsquigarrow$ maximum

The Probit Model

Suppose if we're interested in regression

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description
Assume the following data generation process

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description

Assume the following data generation process

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description

Assume the following data generation process

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(\pi_i) \\
\pi_i &= \Phi(\boldsymbol{X}_i \boldsymbol{\beta})
\end{aligned}
$$

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description

Assume the following data generation process

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= \Phi(\boldsymbol{X}_i\boldsymbol{\beta}) \end{aligned}$$

where

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description

Assume the following data generation process

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= \Phi(\boldsymbol{X}_i\boldsymbol{\beta}) \end{aligned}$$

where

$\Phi(\cdot)$ is the cumulative normal distribution function

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description

Assume the following data generation process

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(\pi_i) \\ \pi_i &= \Phi(\boldsymbol{X}_i \boldsymbol{\beta}) \end{aligned}$$

where

- $\Phi(\cdot)$ is the cumulative normal distribution function
- $\boldsymbol{X}_i = (1, x_i)$

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description

Assume the following data generation process

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(\pi_i) \\
\pi_i &= \Phi(\boldsymbol{X}_i \boldsymbol{\beta})
\end{aligned}
$$

where

$\Phi(\cdot)$ is the cumulative normal distribution function

$\boldsymbol{X}_i = (1, x_i)$

$\boldsymbol{\beta} = (\beta_1, \beta_2) \rightsquigarrow$ parameters.

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description
Assume the following data generation process

$$\begin{aligned} Y_i & \sim \text{Bernoulli}(\pi_i) \\ \pi_i & = \Phi(\boldsymbol{X}_i \boldsymbol{\beta}) \end{aligned}$$

where

- $\Phi(\cdot)$ is the cumulative normal distribution function
- $\boldsymbol{X}_i = (1, x_i)$
- $\boldsymbol{\beta} = (\beta_1, \beta_2) \rightsquigarrow$ parameters.
- N observations

The Probit Model

Suppose if we're interested in regression $\rightsquigarrow$ prediction, classification, description

Assume the following data generation process

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(\pi_i) \\
\pi_i &= \Phi(\boldsymbol{X}_i \boldsymbol{\beta})
\end{aligned}
$$

where

- $\Phi(\cdot)$ is the cumulative normal distribution function
- $\boldsymbol{X}_i = (1, x_i)$
- $\boldsymbol{\beta} = (\beta_1, \beta_2) \rightsquigarrow$ parameters.
- N observations

Implicit (improper) prior

The Probit Model $\rightsquigarrow$ Objective Function

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) & \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}) \\
& \propto \prod_{i=1}^{N} \Phi\left(\boldsymbol{X}_i \boldsymbol{\beta}\right)^{y_i} \left(1 - \Phi(\boldsymbol{X}_i \boldsymbol{\beta})\right)^{1-y_i}
\end{aligned}
$$

The Probit Model$\rightsquigarrow$ Optimization

Theorem

Suppose $f : \Re^K \rightarrow \Re_+$. If $\boldsymbol{x}^*$ maximizes $\log(f(x))$, then $\boldsymbol{x}^*$ maximizes $f(x)$.

$$\log\left(p(\beta|\boldsymbol{y}, \boldsymbol{X})\right) = \sum_{i=1}^{N} \left[y_i \log(\Phi(\boldsymbol{X}_i\beta)) + (1 - y_i)\log(1 - \Phi(\boldsymbol{X}_i\beta))\right] + c$$

Find $\boldsymbol{\beta}^*$ to maximize $\log\left(p(\beta|\boldsymbol{y}, \boldsymbol{X})\right) \rightsquigarrow$ computational method

Computational Optimization Approaches

Analytic (Closed form)$\rightsquigarrow$ Often difficult, impractical, or unavailable

Computational Optimization Approaches

Analytic (Closed form) $\rightsquigarrow$ Often difficult, impractical, or unavailable
Computational

Computational Optimization Approaches

Analytic (Closed form) $\rightsquigarrow$ Often difficult, impractical, or unavailable
Computational $\rightsquigarrow$ iterative algorithm that converges to a solution (hopefully the right one!)

Computational Optimization Approaches

Analytic (Closed form)$\rightsquigarrow$ Often difficult, impractical, or unavailable
Computational$\rightsquigarrow$ iterative algorithm that converges to a solution (hopefully the right one!)

- Methods for optimization:

Computational Optimization Approaches

Analytic (Closed form)$\rightsquigarrow$ Often difficult, impractical, or unavailable

Computational$\rightsquigarrow$ iterative algorithm that converges to a solution (hopefully the right one!)

- Methods for optimization:
  - Newton's method and BFGS

Computational Optimization Approaches

Analytic (Closed form)$\rightsquigarrow$ Often difficult, impractical, or unavailable
Computational$\rightsquigarrow$ iterative algorithm that converges to a solution (hopefully the right one!)

- Methods for optimization:
    - Newton's method and BFGS
    - Gradient descent (ascent)

Computational Optimization Approaches

Analytic (Closed form)$\rightsquigarrow$ Often difficult, impractical, or unavailable
Computational$\rightsquigarrow$ iterative algorithm that converges to a solution (hopefully the right one!)

- Methods for optimization:
    - Newton's method and BFGS
    - Gradient descent (ascent)
    - Expectation Maximization

Computational Optimization Approaches

Analytic (Closed form)$\leadsto$ Often difficult, impractical, or unavailable
Computational$\leadsto$ iterative algorithm that converges to a solution (hopefully the right one!)

- Methods for optimization:
    - Newton's method and BFGS
    - Gradient descent (ascent)
    - Expectation Maximization
    - Genetic Optimization

Computational Optimization Approaches

Analytic (Closed form)$\rightsquigarrow$ Often difficult, impractical, or unavailable
Computational$\rightsquigarrow$ iterative algorithm that converges to a solution (hopefully the right one!)

- Methods for optimization:
  - Newton's method and BFGS
  - Gradient descent (ascent)
  - Expectation Maximization
  - Genetic Optimization
  - Branch and Bound ...

Newton-Raphson Method

Iterative procedure to find a root

Newton-Raphson Method

Iterative procedure to find a root
Often solving for x when $f(x) = 0$ is hard$\rightsquigarrow$ complicated function

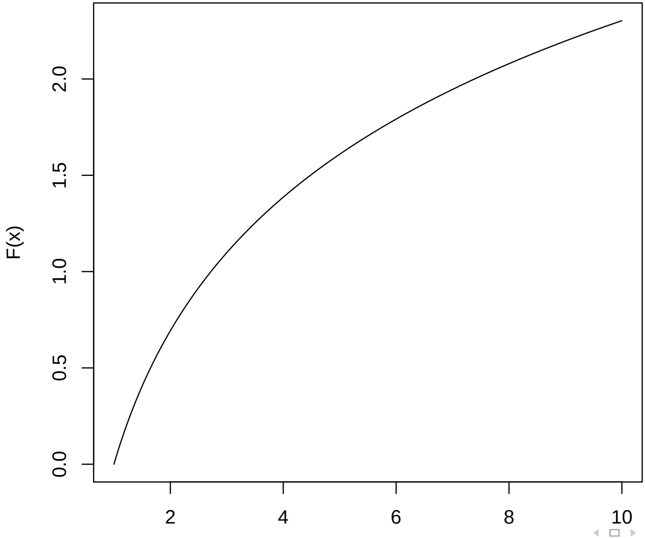
Newton-Raphson Method

Iterative procedure to find a root
Often solving for x when $f(x) = 0$ is hard $\rightsquigarrow$ complicated function
Solving for x when $f(x)$ is linear $\rightsquigarrow$ easy

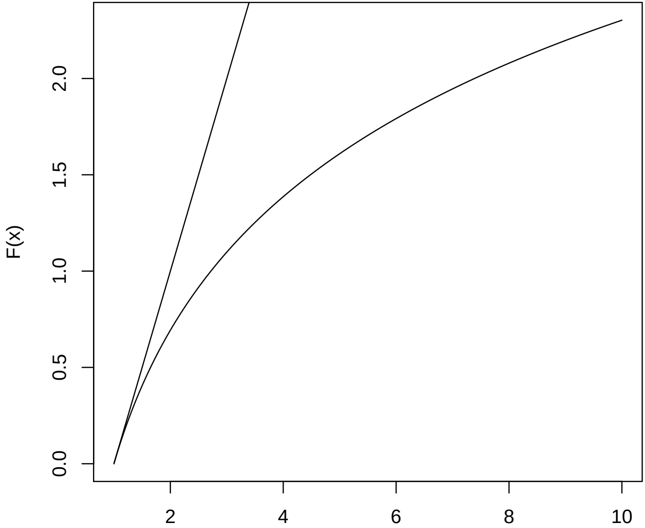
Newton-Raphson Method

Iterative procedure to find a root
Often solving for x when $f(x) = 0$ is hard$\rightsquigarrow$ complicated function
Solving for x when $f(x)$ is linear$\rightsquigarrow$ easy
Approximate with tangent line, iteratively update

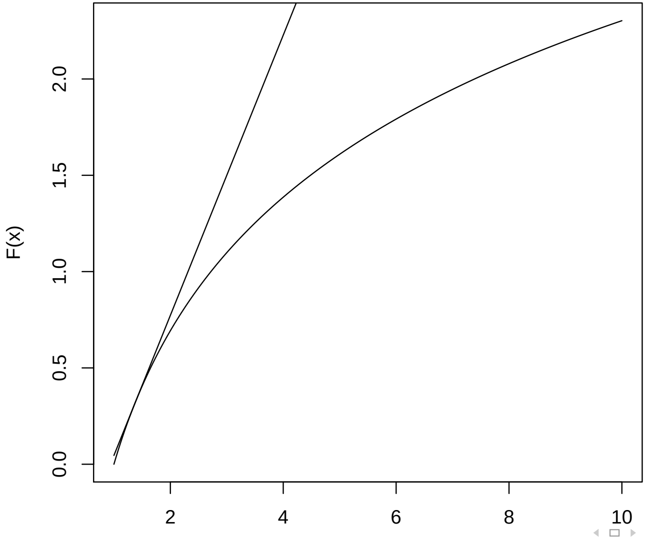
Tangent Line



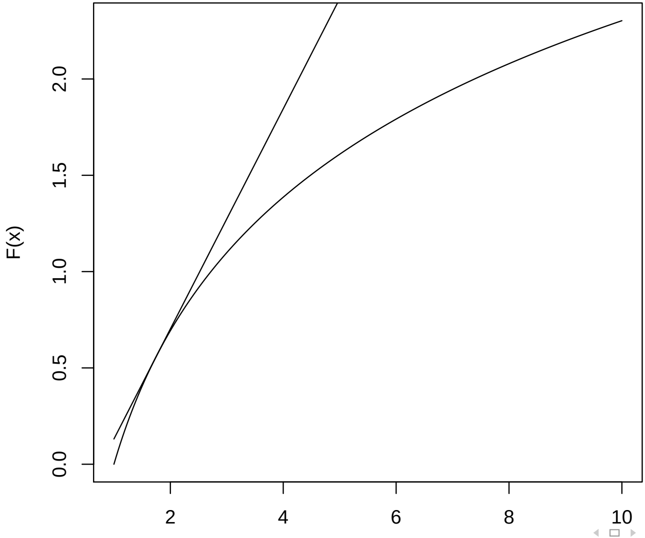
Tangent Line



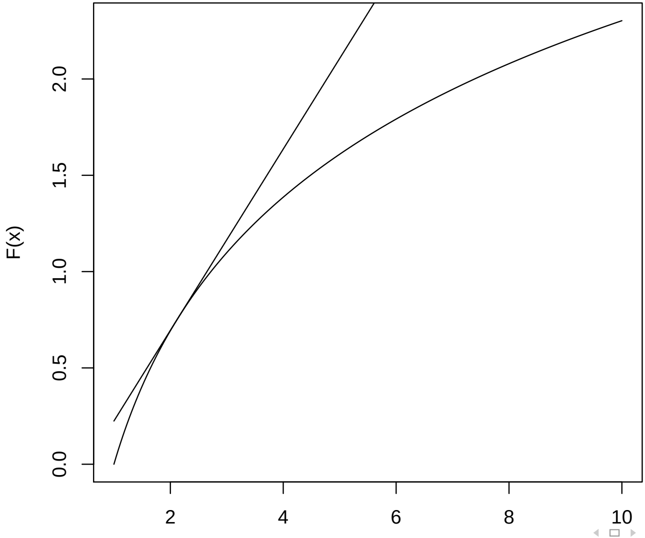
Tangent Line



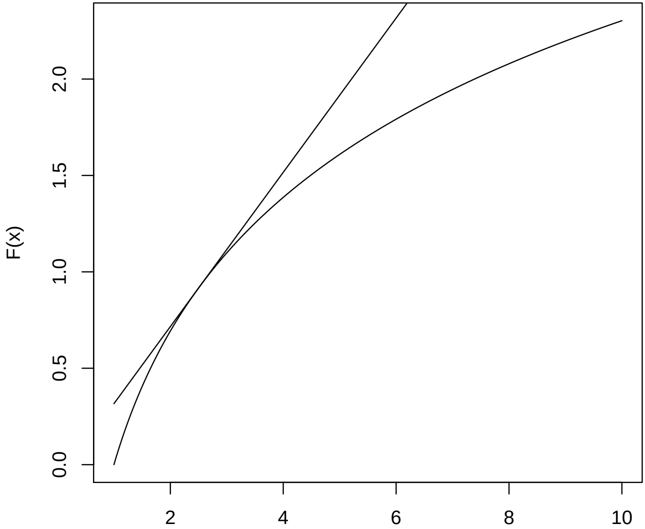
Tangent Line



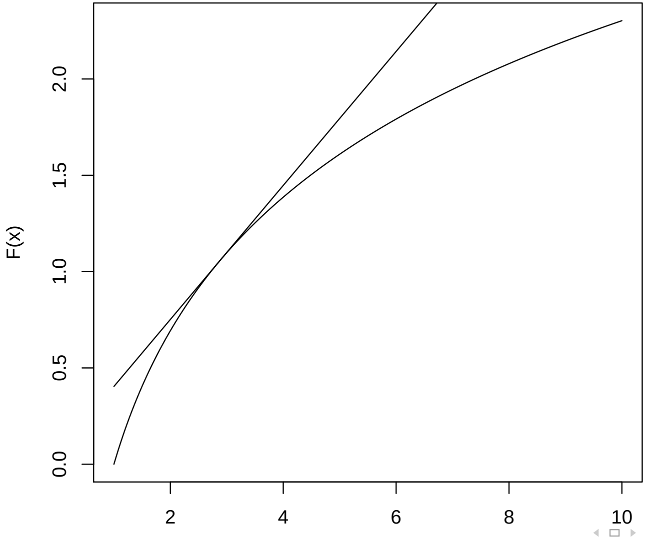
Tangent Line



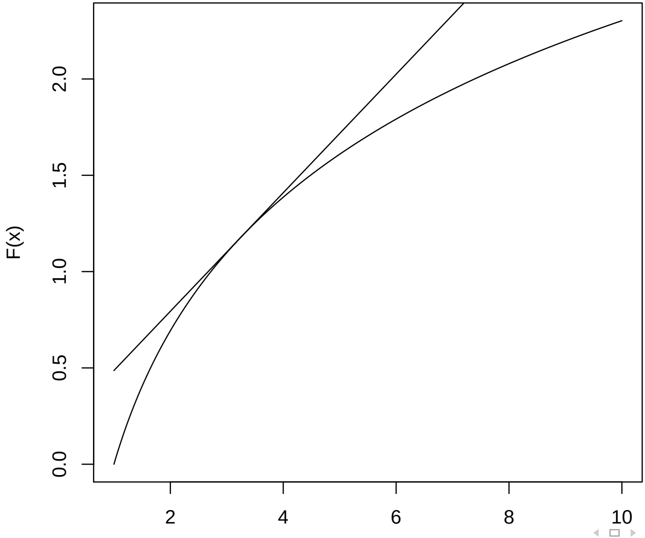
Tangent Line



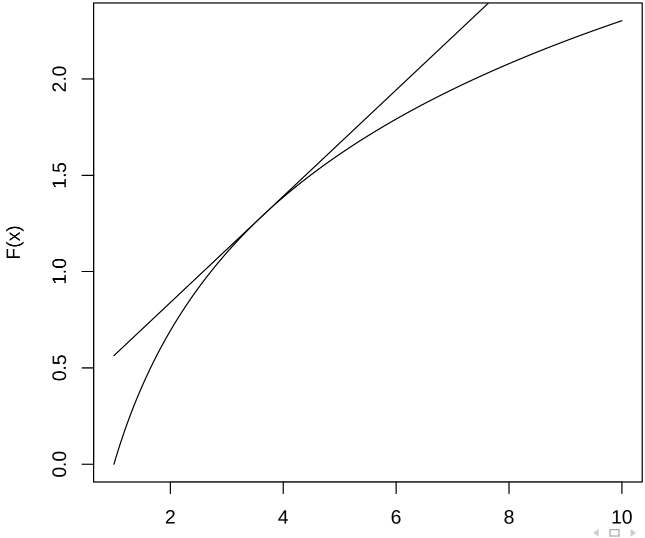
Tangent Line



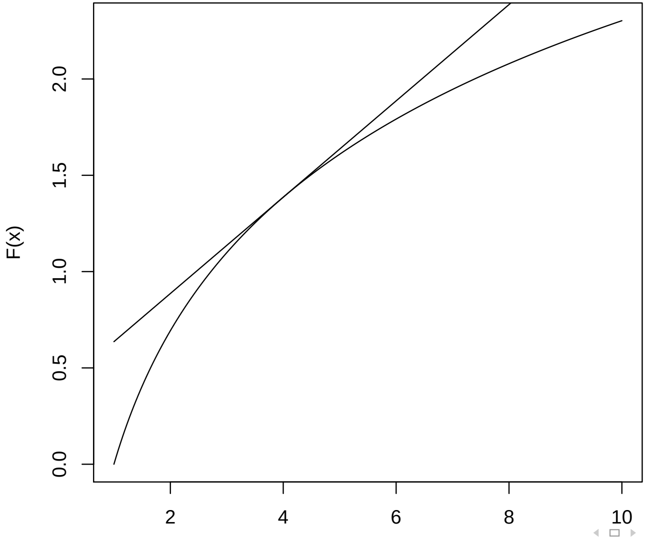
Tangent Line



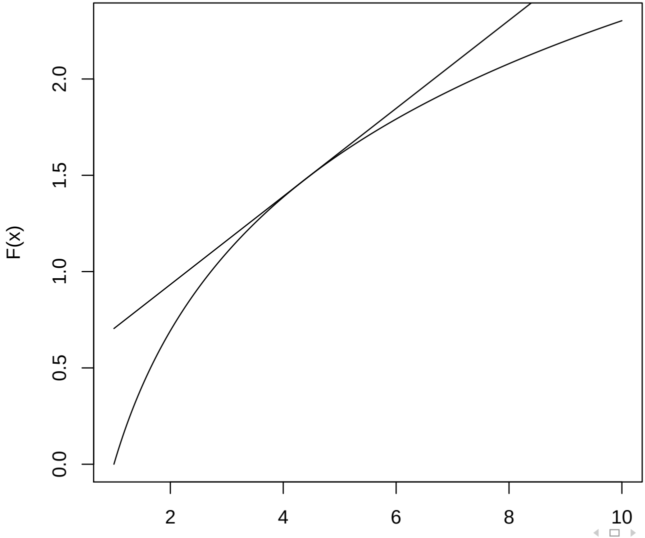
Tangent Line



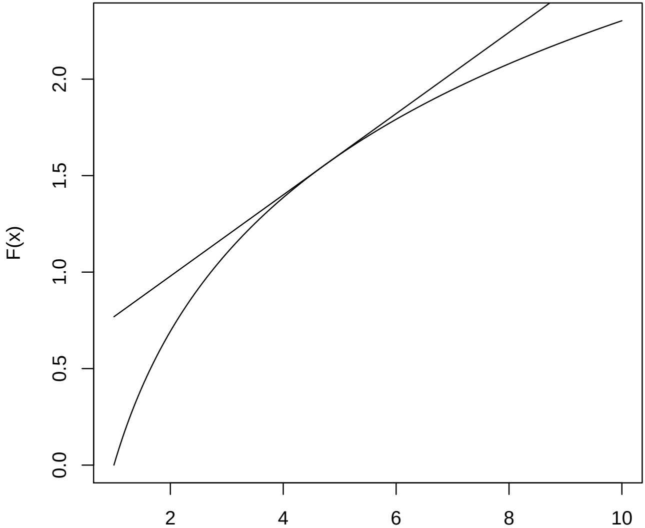
Tangent Line



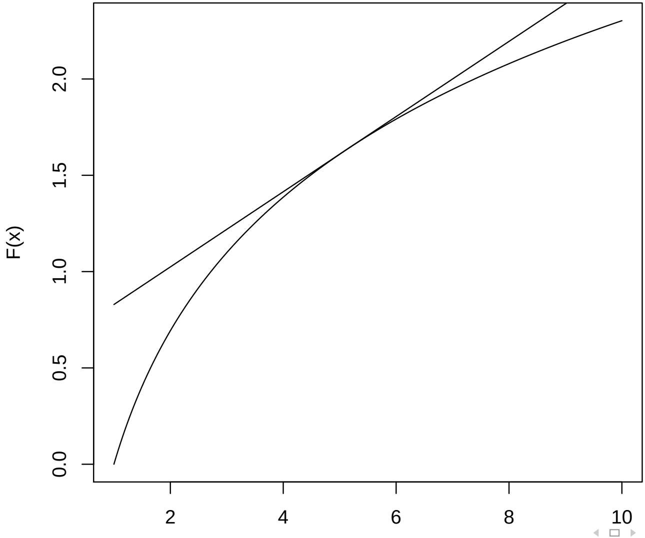
Tangent Line



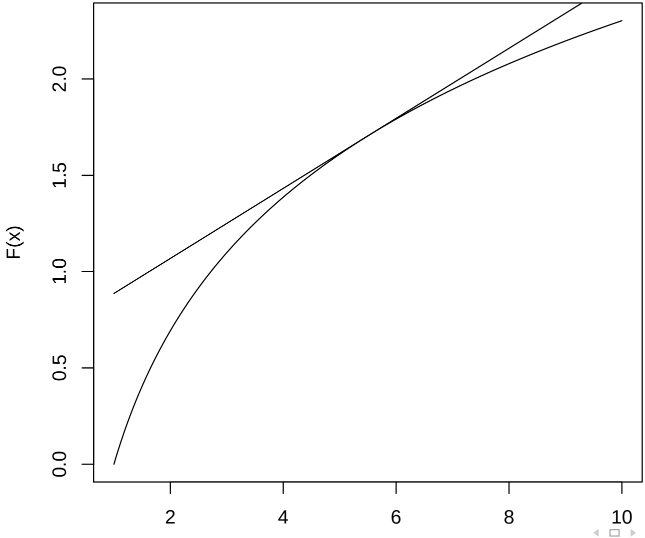
Tangent Line



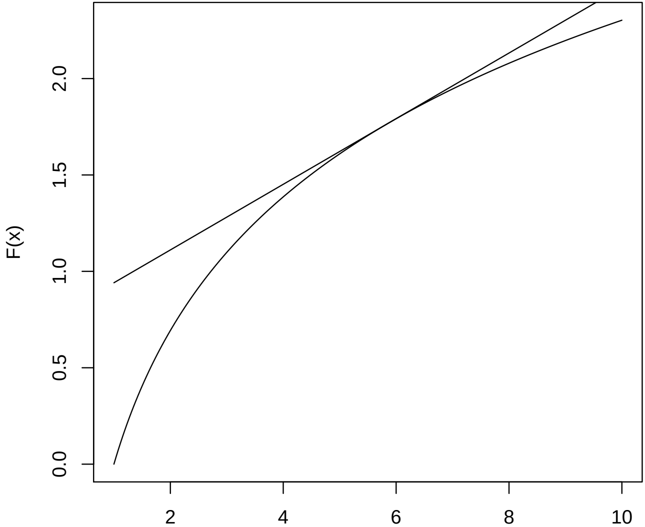
Tangent Line



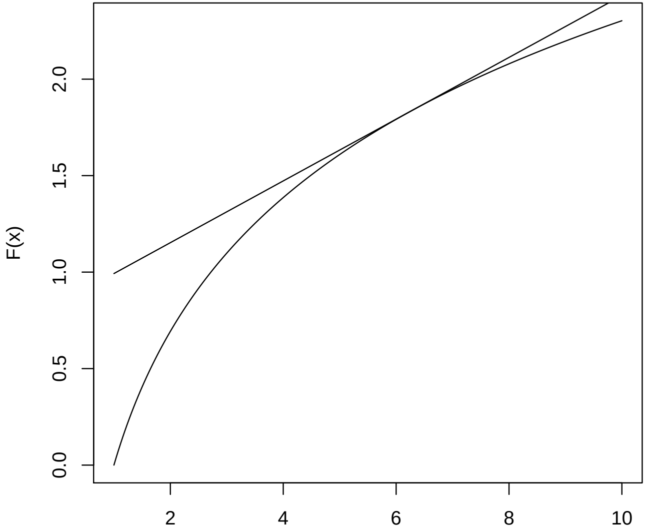
Tangent Line



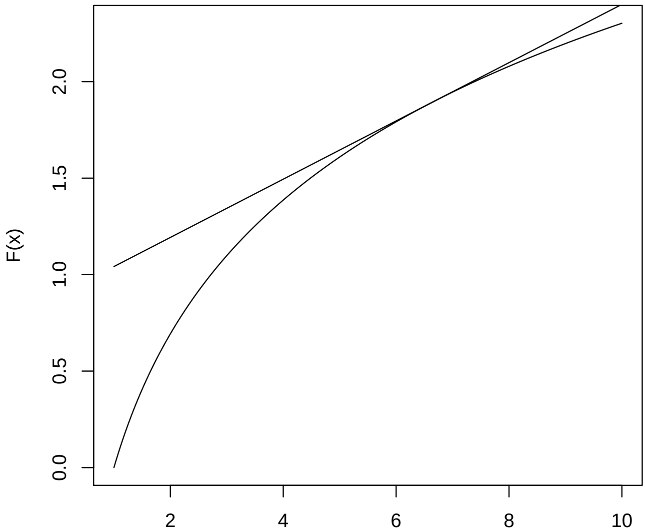
Tangent Line



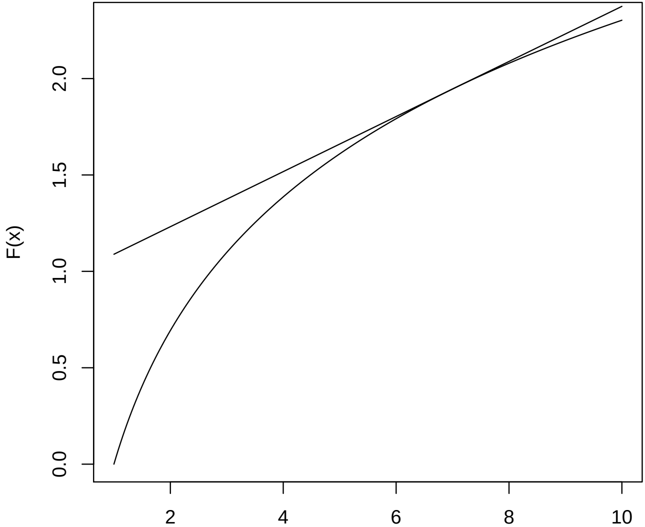
Tangent Line



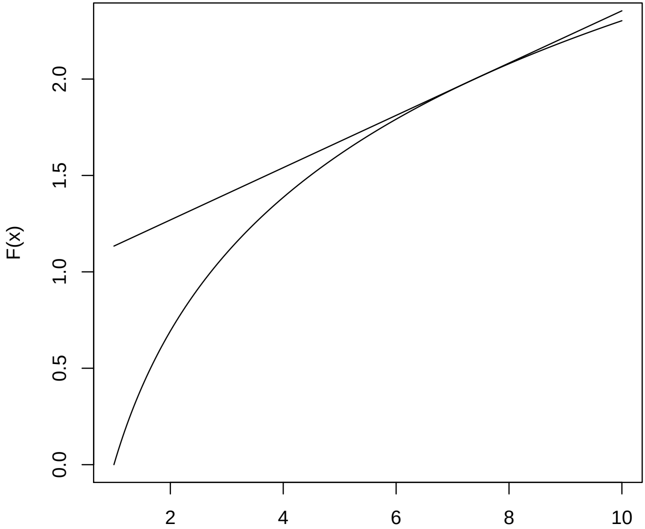
Tangent Line



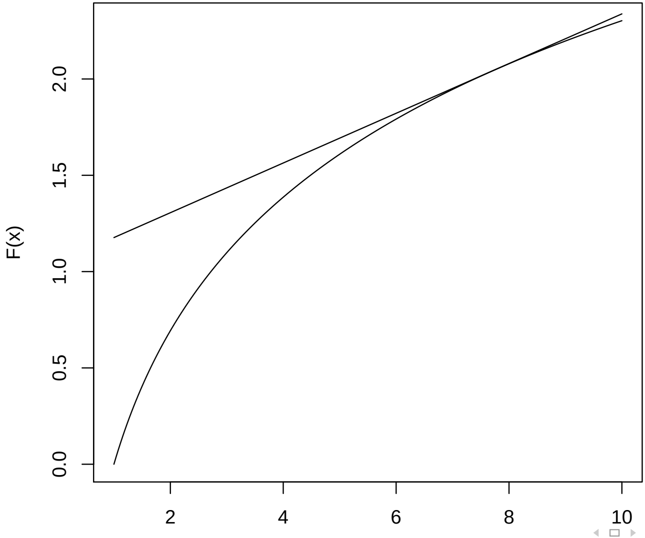
Tangent Line



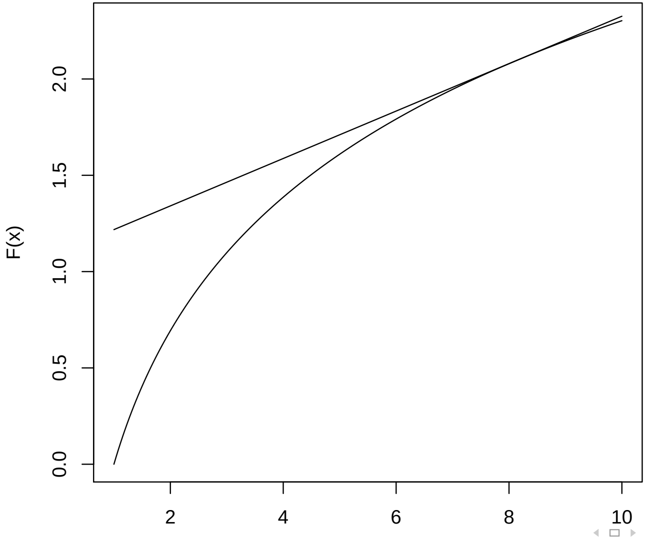
Tangent Line



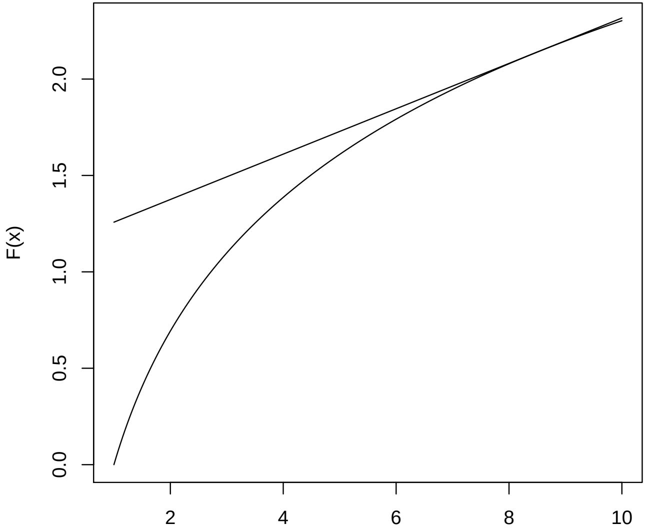
Tangent Line



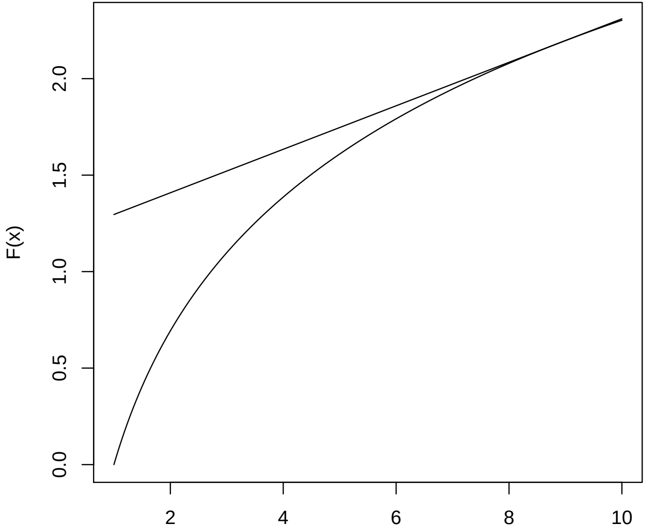
Tangent Line



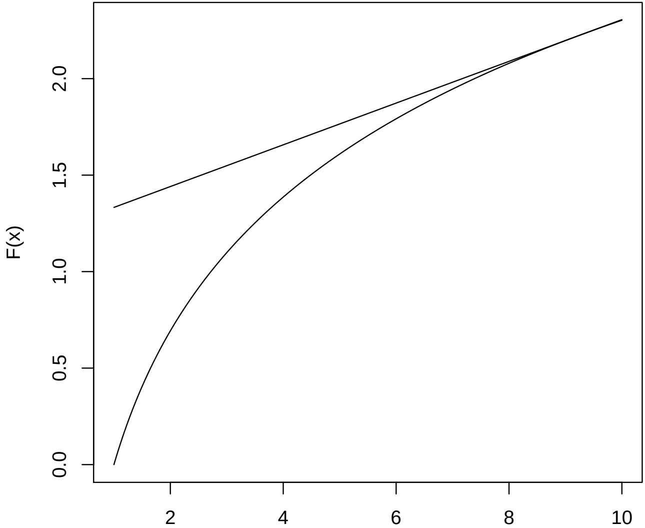
Tangent Line



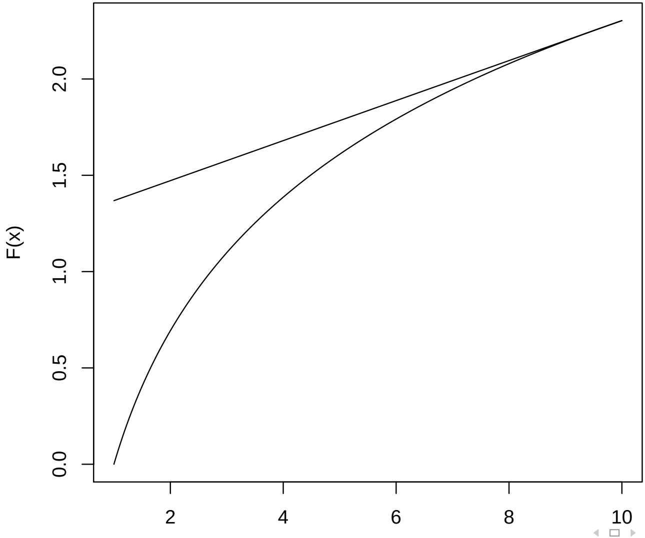
Tangent Line



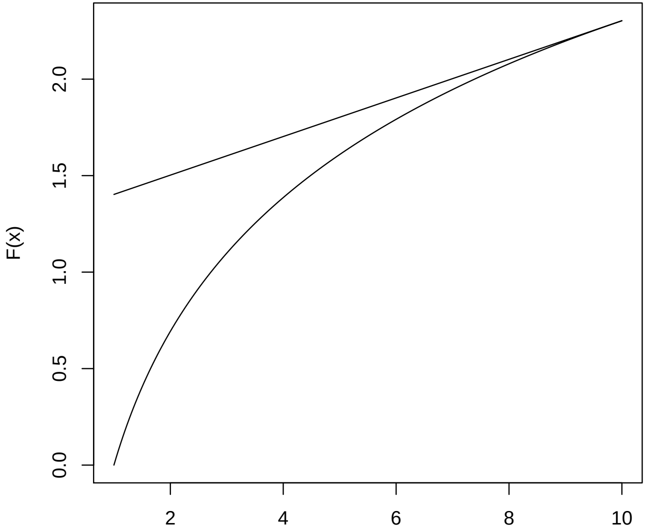
Tangent Line



Tangent Line



Tangent Line



Tangent Line

Formula for Tangent line at x_0:

Tangent Line

Formula for Tangent line at x_0:

$$g(x) = f^{'}(x_0)(x - x_0) + f(x_0)$$

Tangent Line

Formula for Tangent line at x_0:

$$g(x) = f^{'}(x_0)(x - x_0) + f(x_0)$$

Tangent Line

Formula for Tangent line at x_0:

$$g(x) = f^{'}(x_0)(x - x_0) + f(x_0)$$

Tangent Line

Formula for Tangent line at x_0:

$$g(x) = f^{'}(x_0)(x - x_0) + f(x_0)$$

Newton-Raphson Method

Suppose we have some initial guess x_0. We're going to approximate $f^{'}(x)$ with the tangent line to generate a new guess

Newton-Raphson Method

Suppose we have some initial guess x_0. We're going to approximate $f^{'}(x)$ with the tangent line to generate a new guess

$$g(x) \quad = \quad f^{''}(x_0)(x - x_0) + f^{'}(x_0)$$

Newton-Raphson Method

Suppose we have some initial guess x_0. We're going to approximate $f^{'}(x)$ with the tangent line to generate a new guess

$$\begin{aligned} g(x) &= f^{''}(x_0)(x - x_0) + f^{'}(x_0) \\ 0 &= f^{''}(x_0)(x_1 - x_0) + f^{'}(x_0) \end{aligned}$$

Newton-Raphson Method

Suppose we have some initial guess x_0. We're going to approximate $f^{'}(x)$ with the tangent line to generate a new guess

$$\begin{aligned} g(x) &= f^{''}(x_0)(x - x_0) + f^{'}(x_0) \\ 0 &= f^{''}(x_0)(x_1 - x_0) + f^{'}(x_0) \\ x_1 &= x_0 - \frac{f^{'}(x_0)}{f^{''}(x_0)} \end{aligned}$$

Newton-Raphson Method

Suppose we have some initial guess x_0. We're going to approximate $f^{'}(x)$ with the tangent line to generate a new guess

$$
\begin{aligned}
g(x) &= f^{''}(x_0)(x - x_0) + f^{'}(x_0) \\
0 &= f^{''}(x_0)(x_1 - x_0) + f^{'}(x_0) \\
x_1 &= x_0 - \frac{f^{'}(x_0)}{f^{''}(x_0)} \\
x_{t+1} &= x_t - \frac{f^{'}(x_t)}{f^{''}(x_t)}
\end{aligned}
$$

Newton-Raphson Method

Suppose we have some initial guess x_0. We're going to approximate $f^{'}(x)$ with the tangent line to generate a new guess

$$
\begin{aligned}
g(x) &= f^{''}(x_0)(x - x_0) + f^{'}(x_0) \\
0 &= f^{''}(x_0)(x_1 - x_0) + f^{'}(x_0) \\
x_1 &= x_0 - \frac{f^{'}(x_0)}{f^{''}(x_0)} \\
x_{t+1} &= x_t - \frac{f^{'}(x_t)}{f^{''}(x_t)}
\end{aligned}
$$

Perform iteratively until change in $|x_{t+1} - x_t| <$ threshold

Newton-Raphson Method

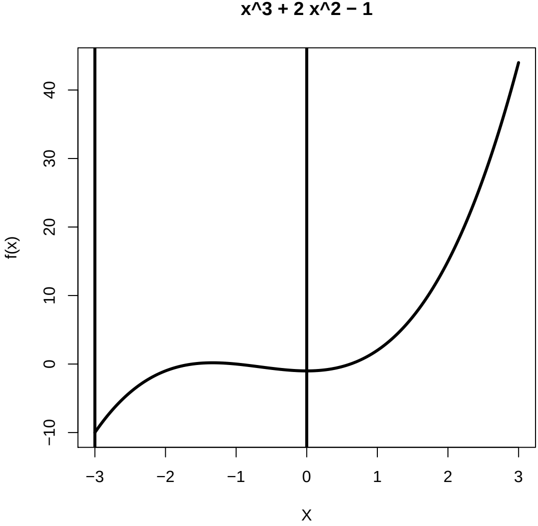
Suppose we have some initial guess x_0. We're going to approximate $f^{'}(x)$ with the tangent line to generate a new guess

$$
\begin{aligned}
g(x) &= f^{''}(x_0)(x - x_0) + f^{'}(x_0) \\
0 &= f^{''}(x_0)(x_1 - x_0) + f^{'}(x_0) \\
x_1 &= x_0 - \frac{f^{'}(x_0)}{f^{''}(x_0)} \\
x_{t+1} &= x_t - \frac{f^{'}(x_t)}{f^{''}(x_t)}
\end{aligned}
$$

Perform iteratively until change in $|x_{t+1} - x_t| <$ threshold

Example Function

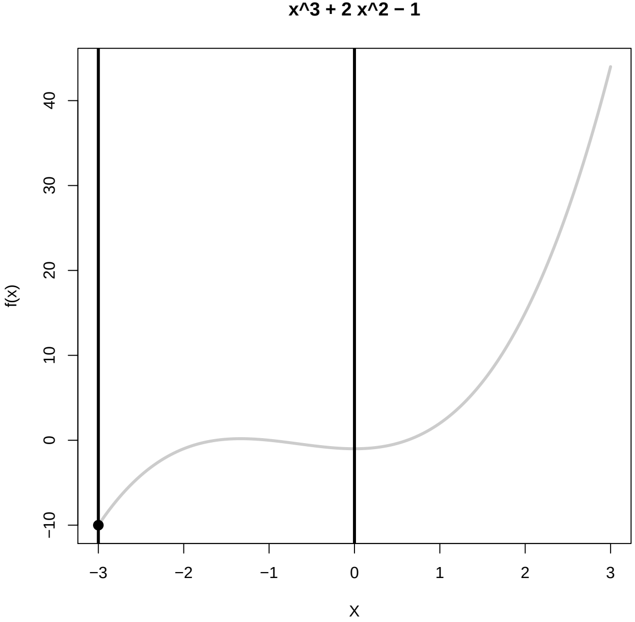
$f(x) = x^3 + 2x^2 - 1$ find x that maximizes $f(x)$ with $x \in [-3, 0]$



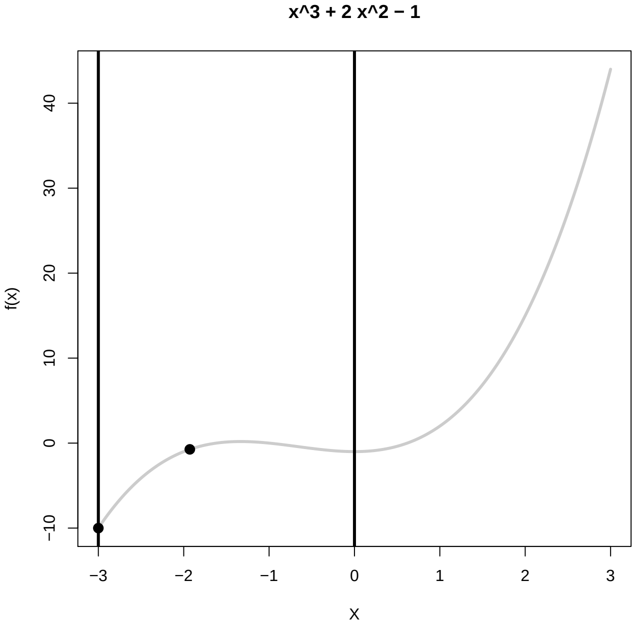
$$
\begin{aligned}
f^{'}(x) &= 3x^2 + 4x \\
f^{''}(x) &= 6x + 4
\end{aligned}
$$

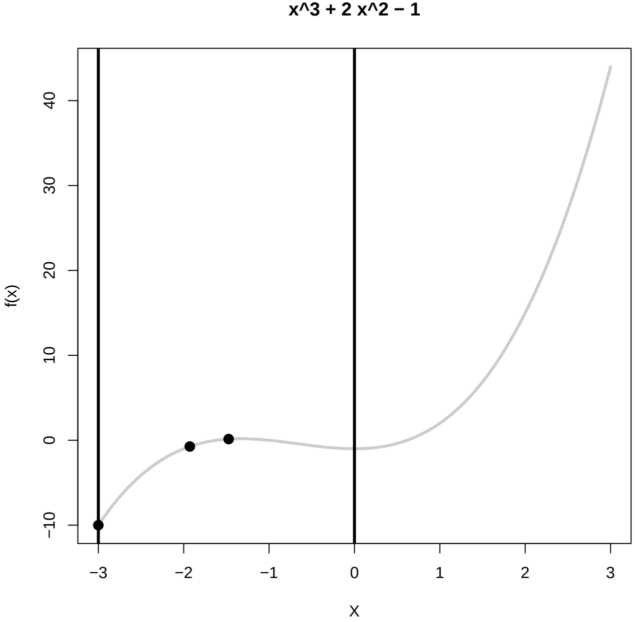
Suppose we have guess x_t then the next step is:

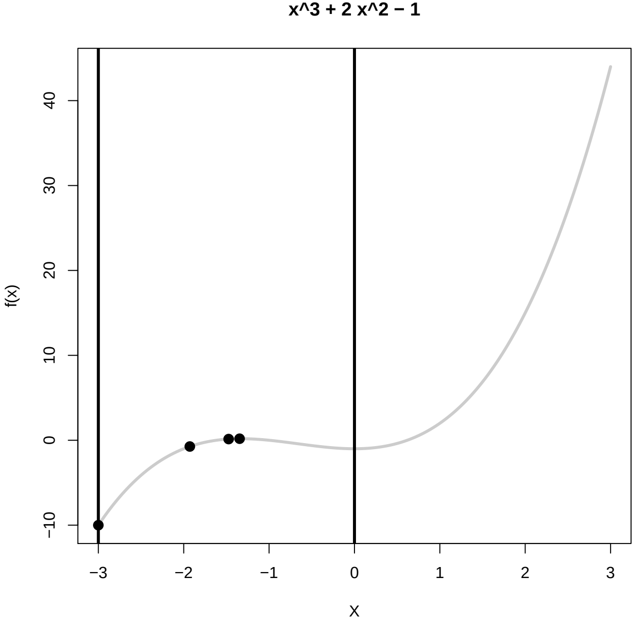
$$
x_{t+1} = x_t - \frac{3x_t^2 + 4x_t}{6x_t + 4}
$$

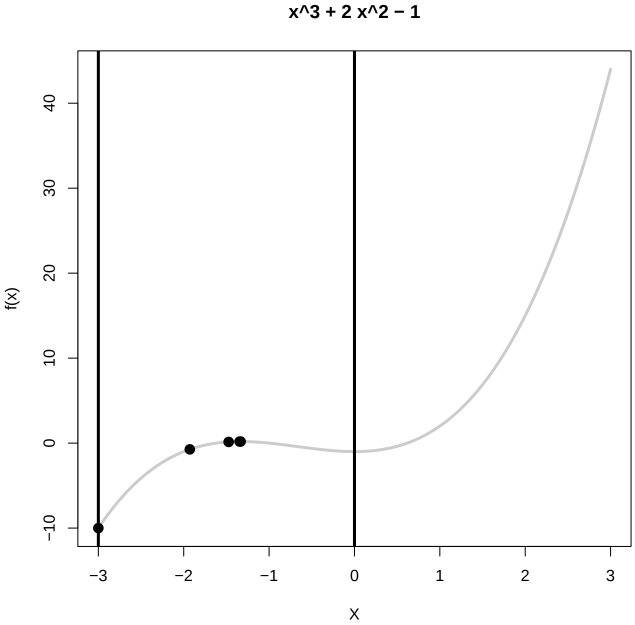


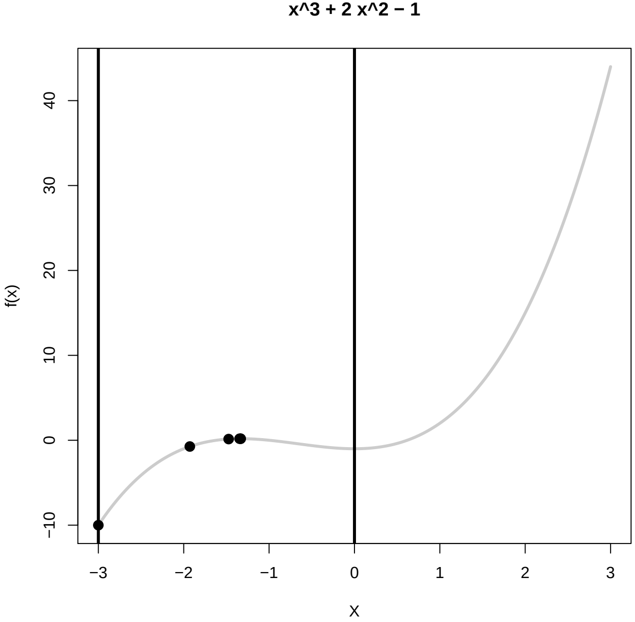
x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X


x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X

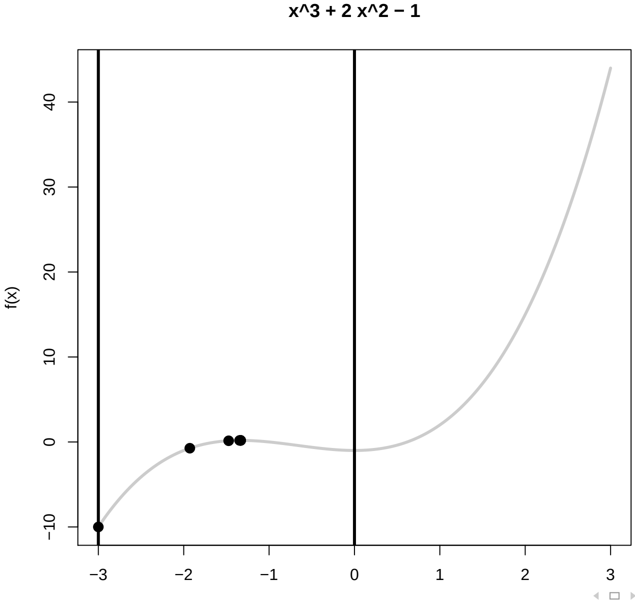

x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X


x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X

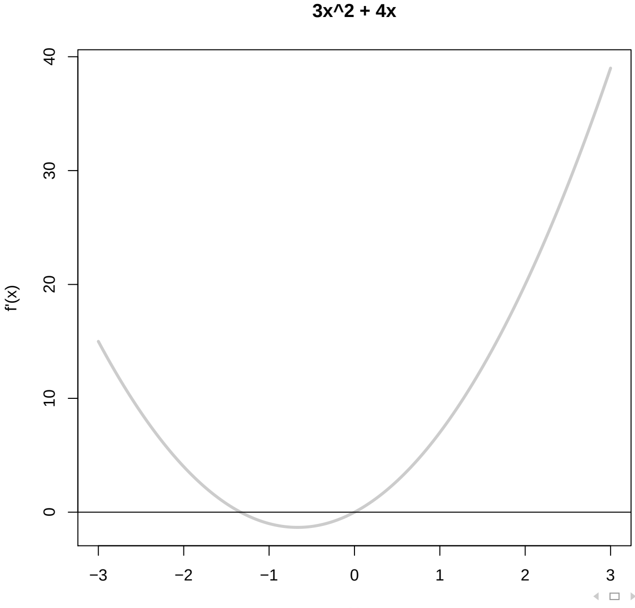

x^3 + 2 x^2 − 1
f(x)
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3
X


x^3 + 2 x^2 − 1
f(x)
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3
X

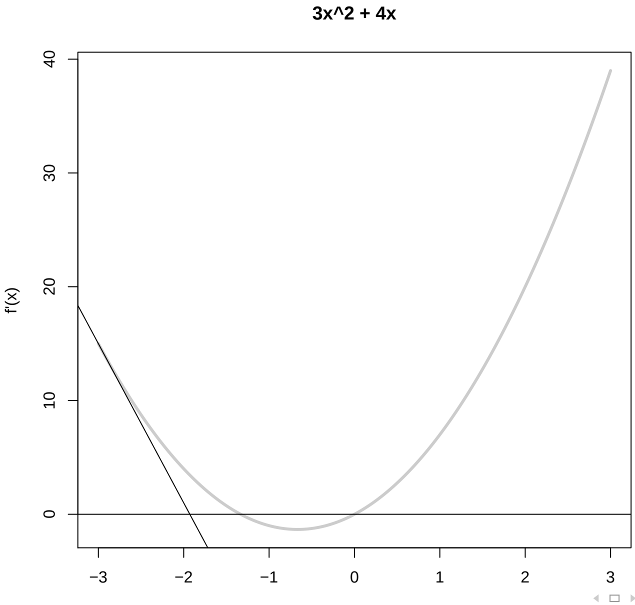
$x^* = -1.3333$



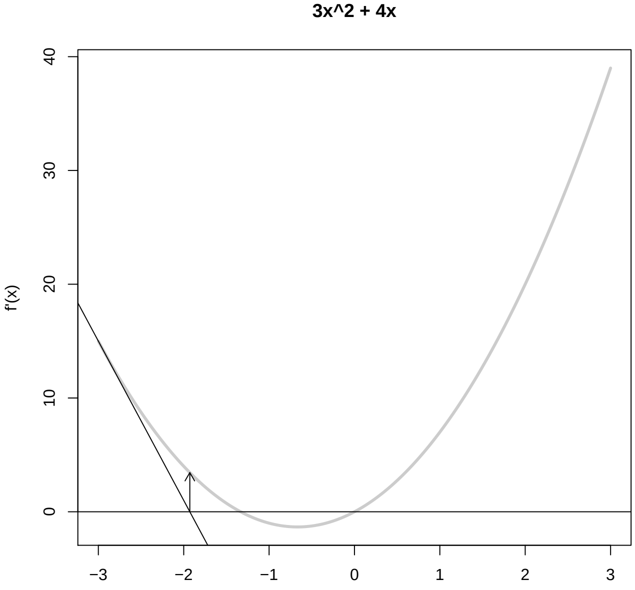
What is Happening with the Roots



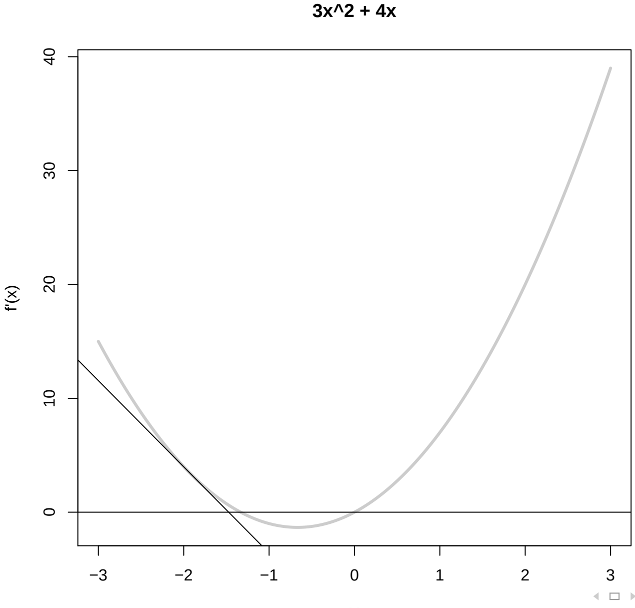
What is Happening with the Roots



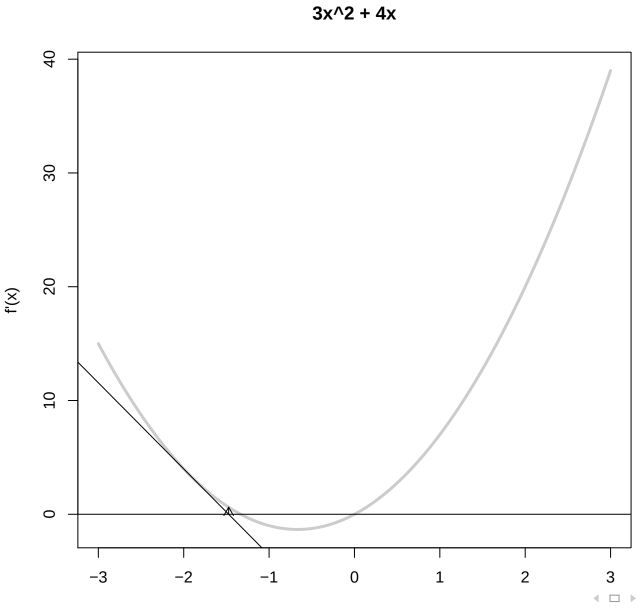
What is Happening with the Roots



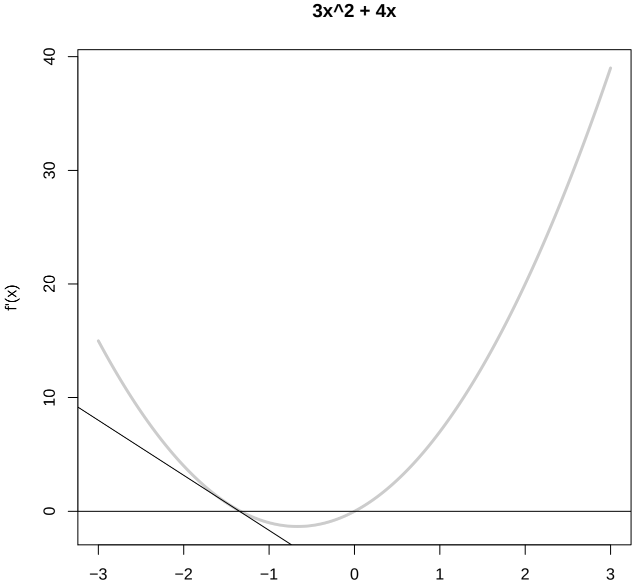
What is Happening with the Roots



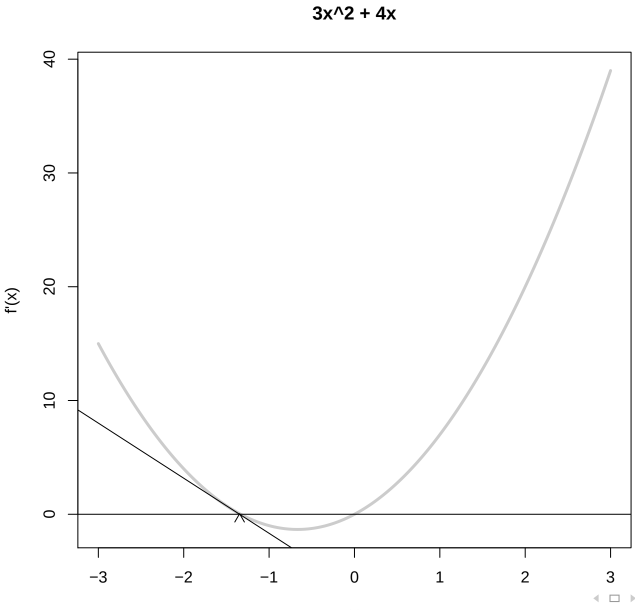
What is Happening with the Roots



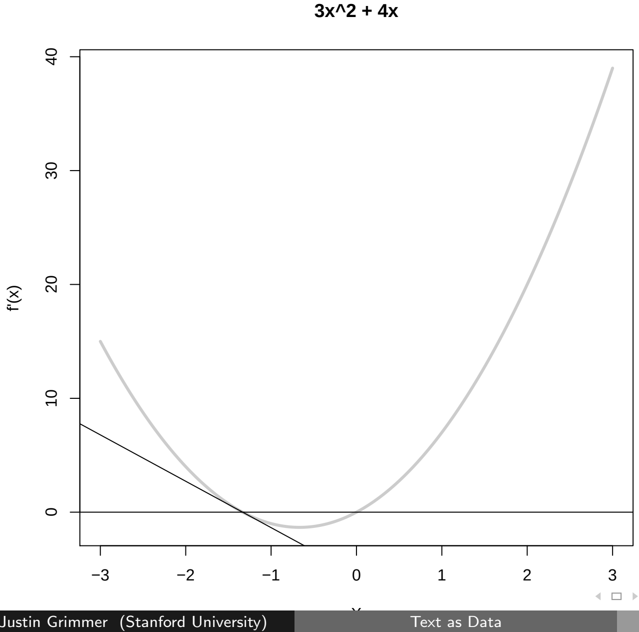
What is Happening with the Roots



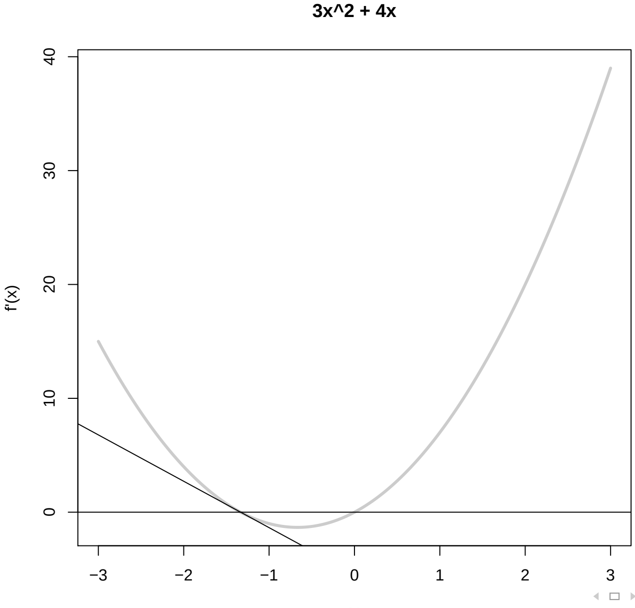
What is Happening with the Roots



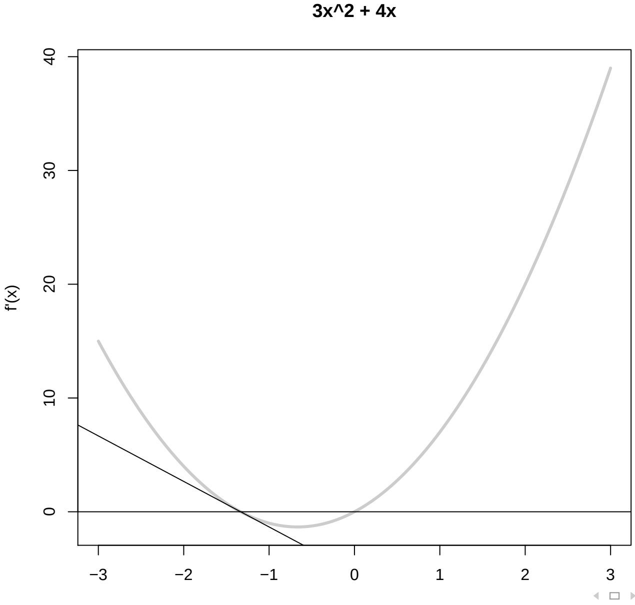
What is Happening with the Roots



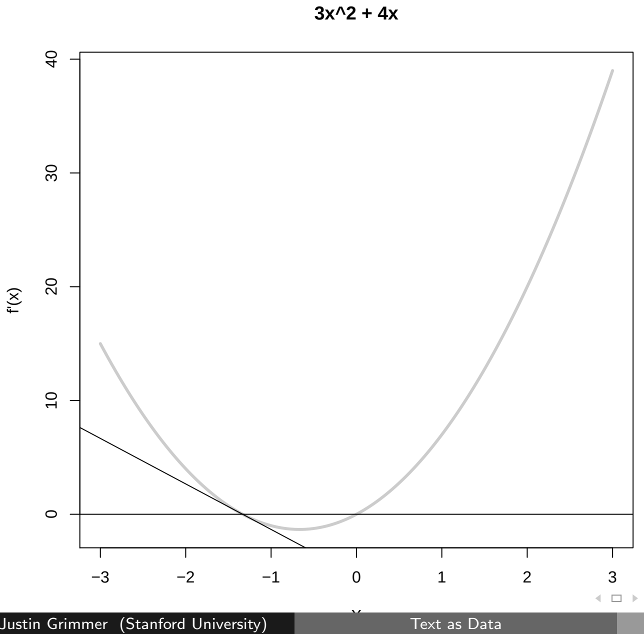
What is Happening with the Roots



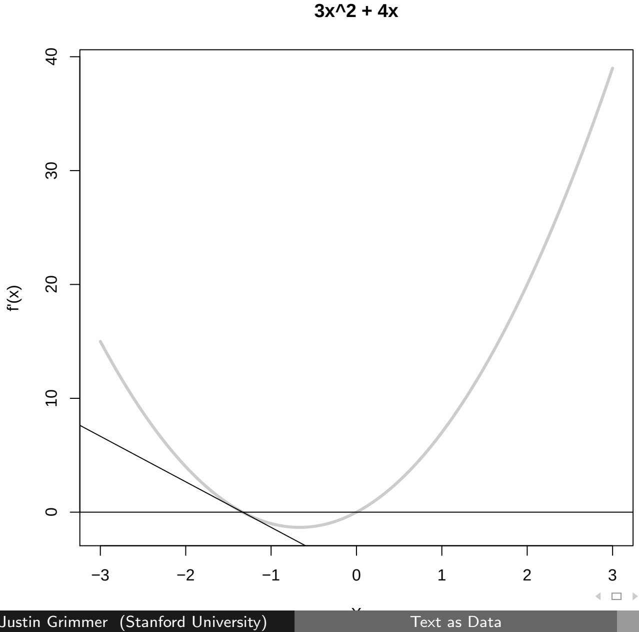
What is Happening with the Roots



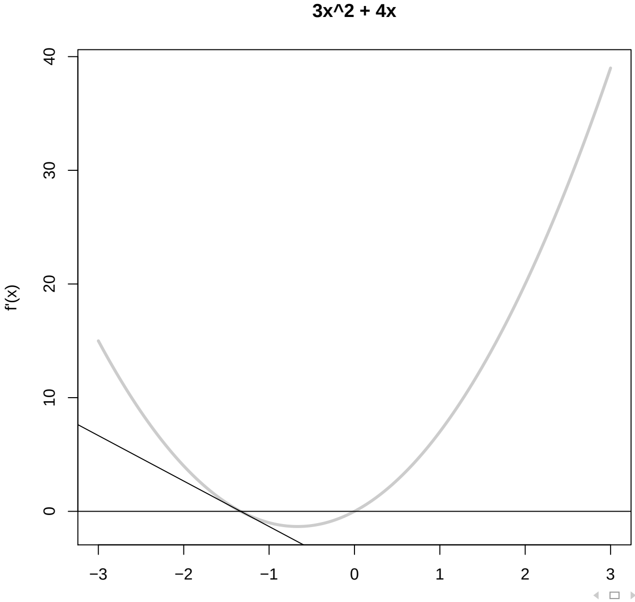
What is Happening with the Roots

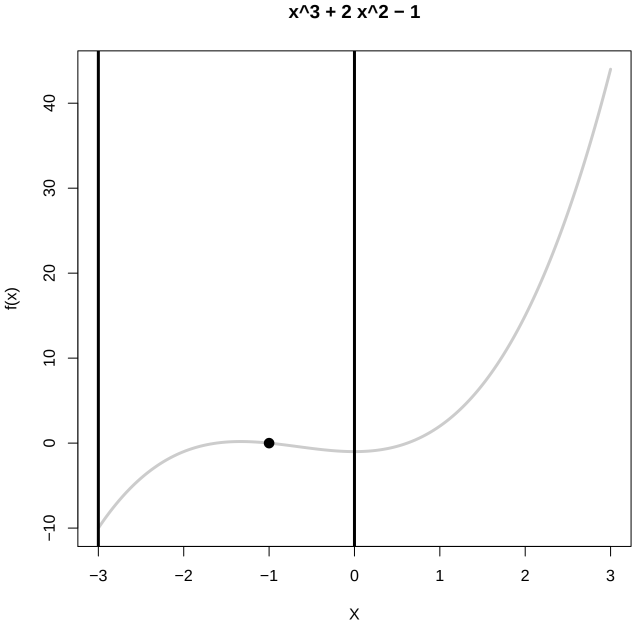


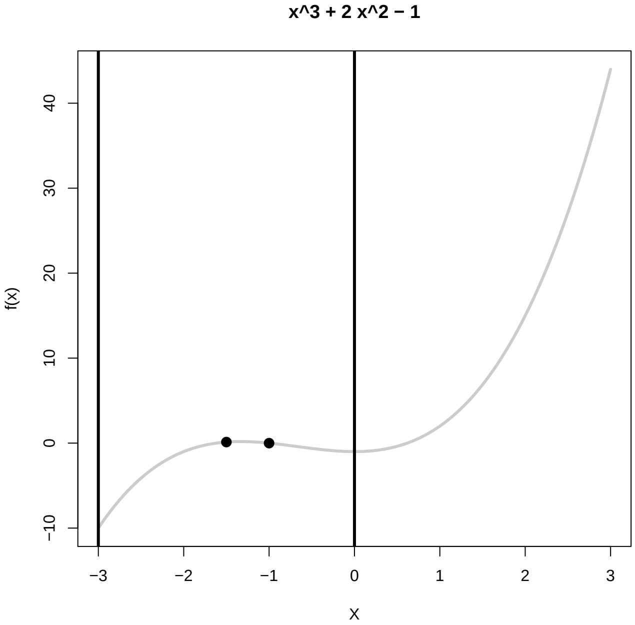
What is Happening with the Roots

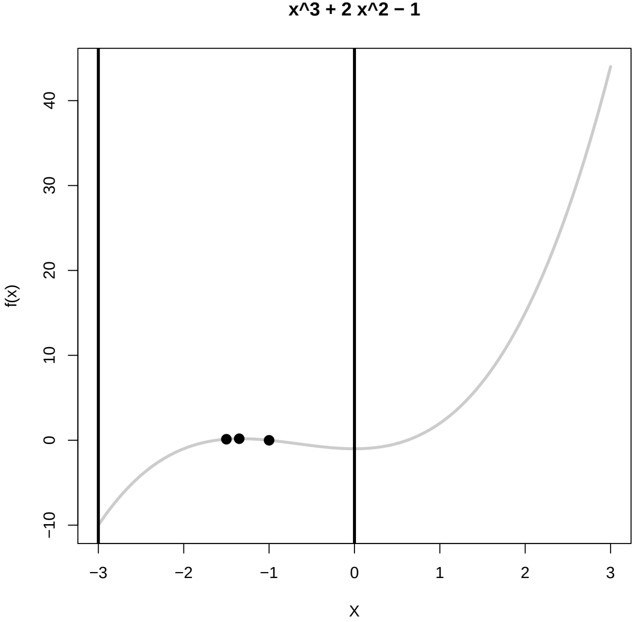


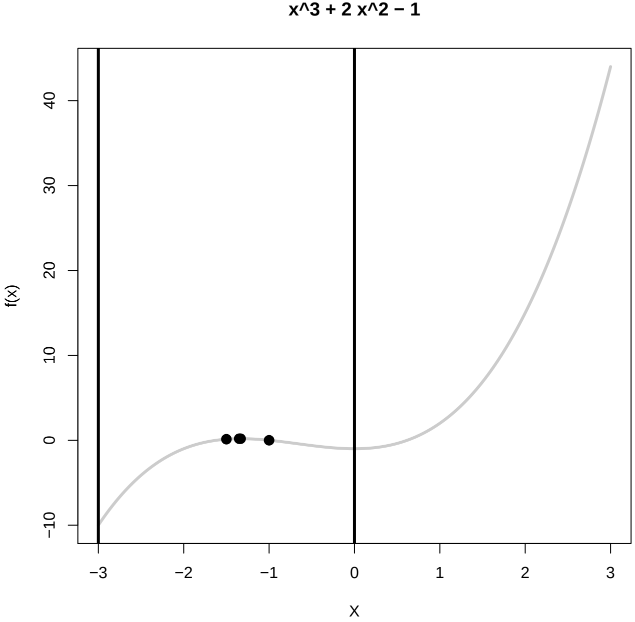
What is Happening with the Roots

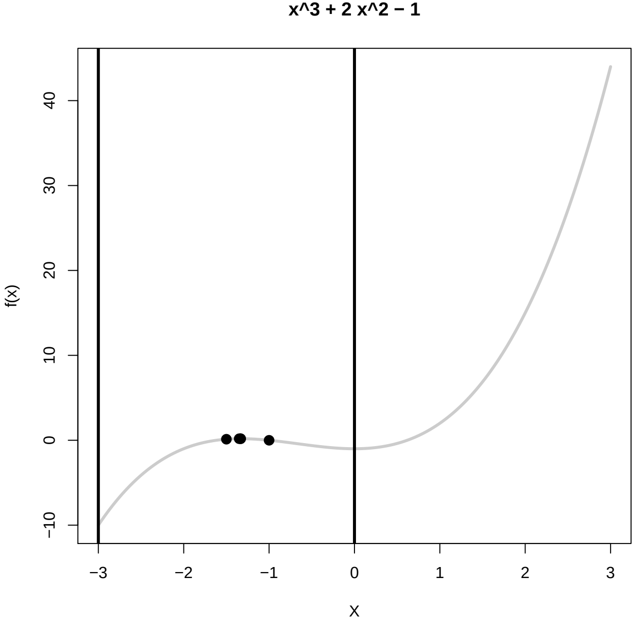


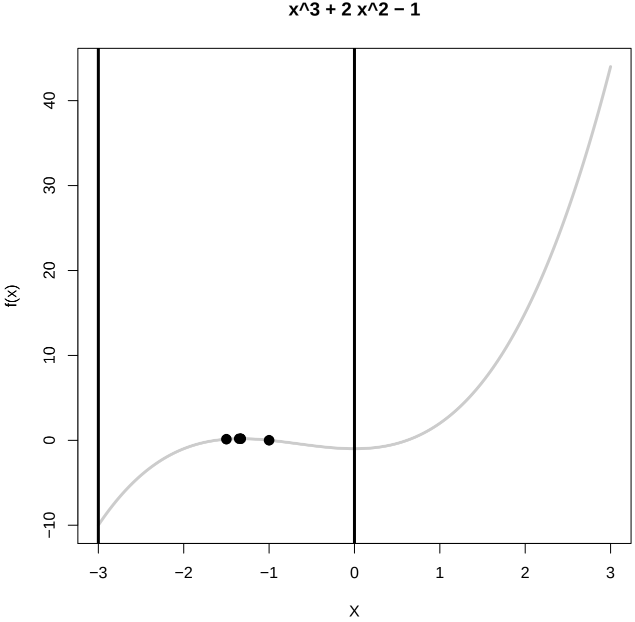

x^3 + 2 x^2 − 1
f(x)
X
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3

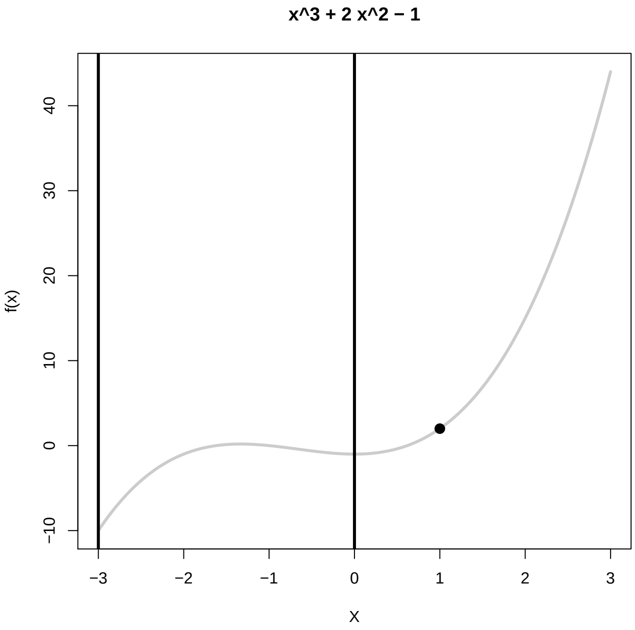

x^3 + 2 x^2 − 1
f(x)
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3
X

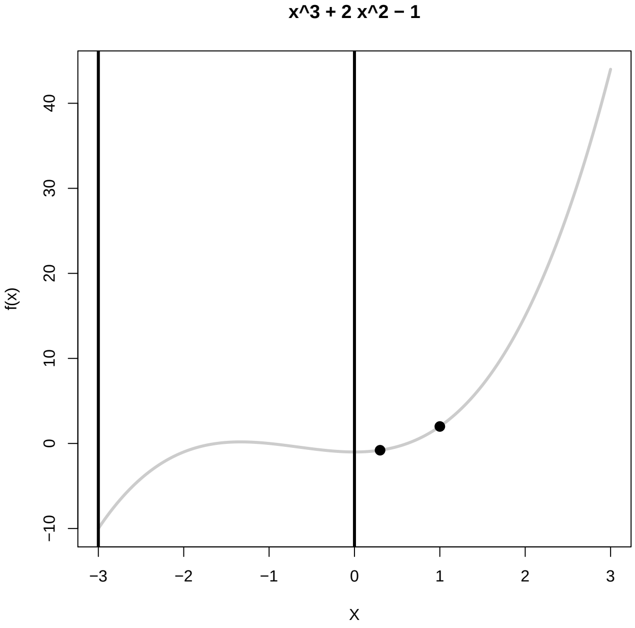

x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X

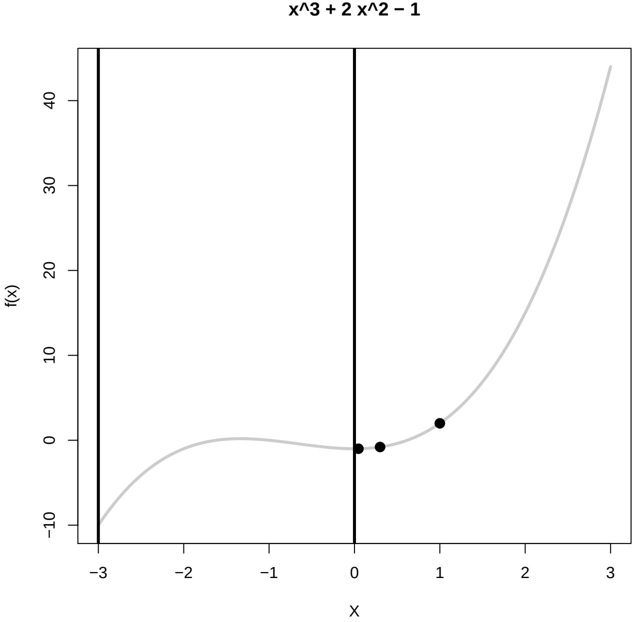

x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X

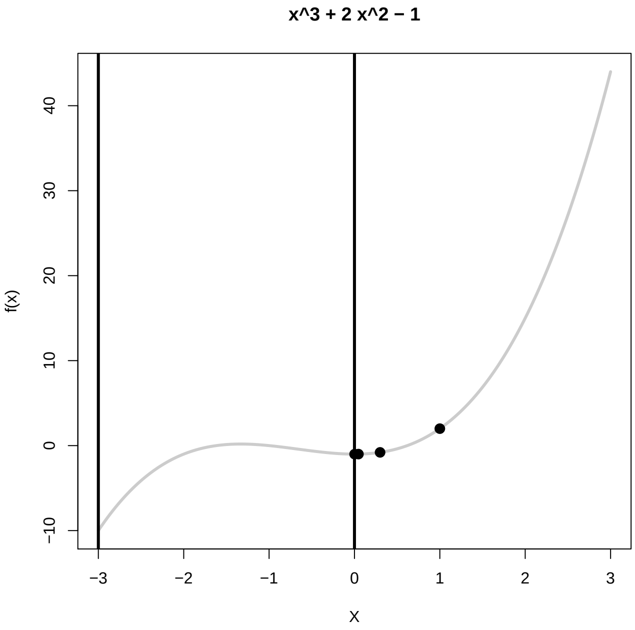

x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X

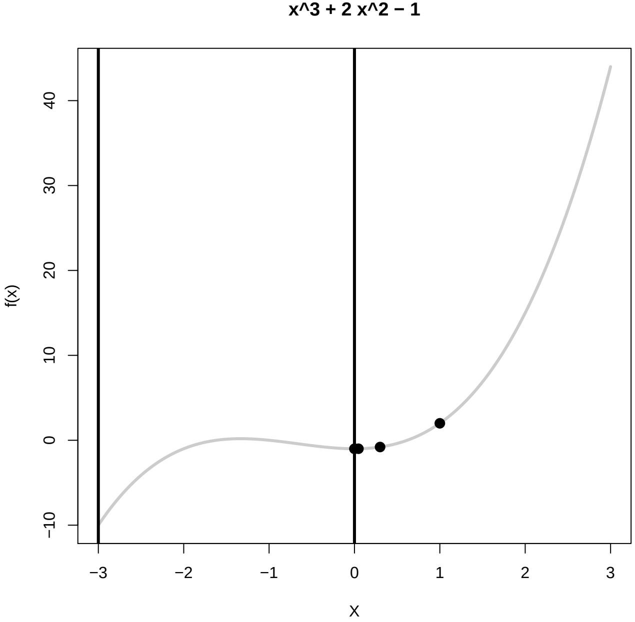

x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X

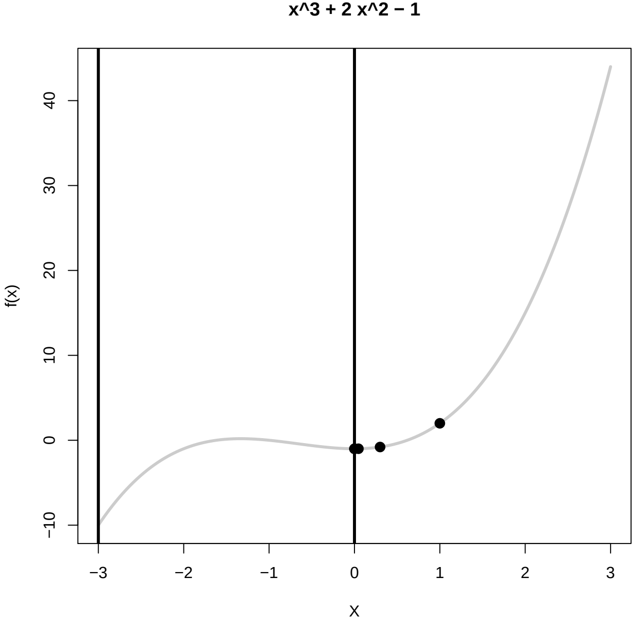

x^3 + 2 x^2 − 1
f(x)
X
−3
−2
−1
0
1
2
3
−10
0
10
20
30
40


x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X


x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X


x^3 + 2 x^2 − 1
f(x)
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3
X


x^3 + 2 x^2 − 1
f(x)
X
−10
0
10
20
30
40
−3
−2
−1
0
1
2
3


x^3 + 2 x^2 − 1
f(x)
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3
X

Multivariate Optimization

$$\log\left(p(\beta|\boldsymbol{y},\boldsymbol{X})\right) = \sum_{i=1}^{N}\left[y_i \log(\Phi(\boldsymbol{X}_i\beta)) + (1-y_i)\log(1-\Phi(\boldsymbol{X}_i\beta))\right] + c$$

To do so:
Apply BFGS (quasi-Newton) in `R`, in `optim`
`R code`

Multivariate Optimization

$$\log\left(p(\beta|\boldsymbol{y}, \boldsymbol{X})\right) \quad = \quad \sum_{i=1}^{N} \left[y_i \log(\Phi(\boldsymbol{X}_i\beta)) + (1-y_i)\log(1-\Phi(\boldsymbol{X}_i\beta))\right] + c$$

To do so:

Apply BFGS (quasi-Newton) in `R`, in `optim`

`R code`

Estimates: predict, classify, describe, ...

Probit Regression, with Prior

Consider the following data generation process

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(\pi_i) \\
\pi_i &= \Phi(\boldsymbol{X}_i \boldsymbol{\beta}) \\
\beta_j &\sim \text{Normal}(\mu, \sigma^2)
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}, \mu, \sigma^2) &\propto p(\boldsymbol{\beta}|\mu, \sigma^2) \times p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}) \\
&\propto \prod_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\beta_j - \mu)^2}{2\sigma^2}\right) \times \prod_{i=1}^{N} \Phi(\boldsymbol{X}_i \boldsymbol{\beta})^{y_i}(1 - \Phi(\boldsymbol{X}_i \boldsymbol{\beta}))
\end{aligned}
$$

Homework $\rightsquigarrow$ explore behavior of $\widehat{\beta}$ as μ, σ^2 vary.

Where We're Going

1) Task
2) Objective Function
3) Optimization procedure

All supposes we have data.
Next week $\rightsquigarrow$ converting text data